Searching or surfing: how do students who use the Web locate information resources?

Viv Cothey
(Vivian John Cothey)
Graduate School of Education
University of Bristol
September 2002

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Social Sciences

61,000

Abstract

This investigation is a large scale study of the real world Web information seeking activity of 1,050 full-time undergraduates studying at a UK higher education institution. The study takes the form of a transaction log analysis of a Web log which records over a two year period all the 1,990,488 urls requested by the students during 46,558 daily sessions. The analysis focuses on how individual students seek Web information. This is made possible by each user being (anonymously) identified throughout the Web log. Both longitudinal and non-longitudinal or repeat study analyses are undertaken. The analyses make use of a novel session-conformance metric which measures the similarity/dissimilarity of the collection of Website requests made during each session.

Over time student-users become more individually distinctive in respect of their 'Web territories' or the collections of Websites which they visit and revisit during each session. Student-users become more territorial in that they increasingly locate their Web information resources from within their own Web territories. 'Searching' occurs in only half of all sessions and student-users undertake less 'searching' as their Web territories become more strongly developed.

These findings are interpreted using the notion of a personal Web information infrastructure which is based on Marchionini's idea of a personal information infrastructure (Marchionini, 1995). A student-user's personal Web information infrastructure is represented by his (or her) territory. As student-users become more proficient at locating Web information resources to satisfy their individual information needs so they build or strengthen their personal Web information infrastructures.

Acknowledgements

I would like to acknowledge all those that have provided me with their invaluable support without some of which this research would not have been possible. A special thanks goes to Rich Woods for his technical wisdom (and beer) and to members of the management team and former colleagues at the University of Gloucestershire for their practical assistance, encouragement and interest.

I would also like to express my particular gratitude to Dr Sally Barnes and Professor Ros Sutherland for their support and advice, and to the UK Economic and Social Research Council for their studentship.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree.

Any views expressed in the dissertation are those of the author and in no way represent those of the University of Bristol.

The dissertation has not been presented to any other University for examination either in the United Kingdom or overseas.

Signed: 

Date: 16 April 2003

Table of contents

List of tables

List of figures

1

How do student-users locate Web information?

1.1 Introduction

The Internet or what has become its major component, the World Wide Web (Web), has leapt to prominence over the last few years (Berners-Lee, 1999) and has captured the imagination not only in a positive sense (Borgman, 2000) but also negatively, for example in connection with so-called cyberporn (Li, 2000). The Web phenomenon spawns much research, for example Jones (1999) but his work is not alone in being criticised for containing "a great number of bold claims ... for which no evidence is provided" (Mathiesen & Fallis, 2000, p. 589). Molyneux & Williams (2001) include the Internet user and use demographics among their measurement categories but introduce their review of "factoids, for-profit sources, systematic studies, and scholarly literature" (2001, p. 289) with:

> The literature of Internet measurement is dispersed, fragmentary, fugitive, and rarely scholarly.
>
> (Molyneux & Williams, 2001, p. 288)

This lack of rigor and scholarlyness only serves to reinforce the motivation which underpins this research which is a sceptical view of existing claims that we know much about Web users, who they are (in a demographic sense) and more importantly what they do. At the outset of the study it seems that there is much froth and little substance. An apparent paradox with the Web is that there is such a vast volume of raw data flowing between computers and being counted (Cooperative Association for Internet Data Analysis, 2002) while at the same time an almost complete absence of reliable use and user information. The most prominent user survey, GVU, is self-selecting (Georgia Tech Research Corporation, 1999). Other periodic user surveys

include *The UCLA Internet report* (UCLA Center for Communication Policy, 2002) and the *Index of Internet connectivity* (ONS, 2002) which although more rigorous are of their nature unspecific. The situation is neatly summed up by Molyneux & Williams who conclude that "Generally, however, the Internet data environment is not friendly toward scholarship" (2001, p. 325).

Discussion within the Web user characterization group (W3C Web characterization activity, 1999) shows also that (not surprisingly) there is a conflict of terminology and emphasis between what seems to be the technology-centric or *usage* perspective adopted by computer scientists and the user-centric or *user* perspective which might be adopted by a library or information scientist (Baeza-Yates & Ribeiro-Neto, 1999). The difference between the user and usage points of view is crucial as is illustrated by the following hypothetical example which concerns a *transaction log analysis* (TLA) based investigation of an *online public access catalogue* (OPAC).

Suppose men always use two query terms when searching for an item while women always use three; both men and women users each submit the same number of search queries. The library management wish to evaluate the implementation of an OPAC education programme designed to encourage the use of more query terms and therefore undertake a pre and post TLA. This reveals that out of a thousand searches the average number of query terms has increased from 2.4 terms per query to 2.6 terms per query (which is statistically significant). The programme is therefore judged a success.

However, unknown to the library the pre and post gender proportion of OPAC users is different and moved from 60:40 in favour of men to 60:40 in favour of women. Hence the analysis of OPAC searches is,

$$\text{pre-programme: } \frac{600 \times 2 + 400 \times 3}{1,000} = 2.4 \text{ terms per query}$$

and

$$\text{post-programme: } \frac{400 \times 2 + 600 \times 3}{1,000} = 2.6 \text{ terms per query}$$

Thus the increase in the number of terms per query is entirely explained by the demographic change in the user population. The library management validly studied *usage* but their study of *users* is flawed. In this example, if men and women each submit the same number of terms in search queries then arithmetically the study would produce a *correct* answer. Methodologically the study remains flawed but its fallibility is only material if the population is heterogeneous with respect to the metric under consideration.

Because how individuals use the Web is not known, the effect of variation among individuals is not known. Therefore existing findings about using the Web are intrinsically fallible.

It will be found also that Web research suffers from the lack of a consistent terminology and either ignorance or lack of rigor by investigators. Two areas of difficulty demonstrate the point. Firstly, what is meant by the description Web *site*? Two distinct usages are (a) the portal site of a service or some other commercial enterprise *inclusive* of all the resources apparently *contained* within that site, and second it may mean (b) an individual resource. So Web site can have either a collective or particular meaning which makes it difficult to carry out use analysis. As will be seen, equating Web site with Web *page* or *uniform resource locator* (url) is an improvement but is an incomplete solution.

The second example concerns *embedded* image files (but could equally be about, for example, *caching* or *Internet Protocol* (IP) addresses). Web research imposes an obligation on the researcher to acquire a level of knowledge about Web mechanisms sufficient to reliably and validly interpret the research data. But it will be found that much research is flawed, possibly fatally by failures of knowledge. For example, the BBC under the headline *UK Web stats "notoriously inaccurate"* quotes the head of an international research company and says that "firms do not understand the technical complexities ... and so misrepresent their statistics unknowingly" (BBC, 2001).

There is no clear definition of terminology to describe a user's interaction with a hypertext such as the Web (Pitkow, 1997) and the lack of common metrics makes it difficult if not impossible to compare arithmetically the results of different studies. A spatial metaphor is commonly used, thus one *navigates* the Web by *jumping* from site to site as one *visits* each in turn. *Clicking* as in clicking on or visiting a Website as well as *requesting* information from a Website is also used. Hence one can count *clicks* and analyse the *clickstream* or sequence of Website visit requests (Berners-Lee, 1999). One might also *browse*, *surf* or *search* the Web although *searching* the Web frequently refers restrictively to using an *information retrieval* (IR) type Web *search-engine* (Hsieh-Yee, 2001).

The terminology and metrics which are used in this dissertation are developed and defined as required in order to be as rigorous as is needed by the thesis. (The terminology is described in the Glossary.) In particular the terminology *Website* and *Webhost* are defined. The definition of a Website is linked to information. The intent of the definition is that if two users each request to view the same file then, from the point of view of this research, they will have visited the same Website. The number of such requests is measured in clicks.

The research here therefore sets out to separate some of the myth from the measured. It aims to discover what it is that a large group of student-users of the Web *really* do. Several distinct approaches to this research might be taken each of which reflects a different interpretation of the meaning of the question *what do users really do*? These distinct approaches would also reflect the perspective of the investigator. Web information, usage, interaction etcetera are all validly (but differently) investigated by computer scientists, psychologists, mathematicians, artists, specialists in marketing as well as by information scientists.

The research which is described here reflects the user-centric information science inclination of the investigator. The research interprets *how users locate Web information* in the sense of *what it is that users do*, or, *what are users' information seeking actions*? These Web information seeking actions are captured by carrying out an unobtrusive *Web log* survey of the time and url of each Website which each user visits. The method used is in no way *cognitive* (Ingwersen, 2001; Wang, 2001) and no attempt is made to describe the process by which a user constructs meaning from Web information, nor the process by which a *need* for Web information is *satisfied* (for example, Ellis, 1992; Wilson, 1999).

The investigation comprises two extended Web log surveys. The first study-year lasts 284 days and the second study-year lasts for 309 days. Since the two study-years are analysed separately but consistently then the research forms a repeat study. This allows conclusions to be more rigorously examined by comparing the study-years. Individual users are identified anonymously but consistently throughout the entire two year period. Hence, in addition, how an individual student-user locates Web information while less experienced can be compared with how that individual student-user locates Web information when more experienced. This longitudinal design feature is specifically endorsed by Mayer who concludes his discussion of the methodological problems of novice and expert computer user research by drawing attention to longitudinal studies which "offer an excellent complement" (1991, p. 578) to more traditional novice-expert designs.

The lengthy overall duration of the survey period exposes the study to being confounded by both changes in Web information seeking behaviour which are individual, for example as users become more proficient, and by changes in the structure of the Web. These *structural* changes could affect both the mechanisms relating to Web information seeking activity and the distribution of information within the Web. Hence such changes in the Web may provide alternative explanations for longitudinal phenomena in Web information seeking activity.

The Web log is a rich data source and can be analysed and interrogated endlessly. For example analyses can be undertaken focusing on either the user or the *session* where

a user's session is usually thought of as being a collection of transactions which occur close together in time and relate to the same information need. It will be seen that both theoretically and practically the notion of session is problematic. The research presented here takes a pragmatic approach and considers a user's Web information seeking session as being all the Web information seeking activity undertaken during a single day. Analyses which focus on the user and where the Web log data for each user is separated are called *by-user* analyses. In *by-session* analyses the distinction between different users in the Web log data is lost while in a *by-click* analysis the distinction between different sessions is also lost.

The next Section describes the institutional background and context of the survey. The penultimate Section describes how the dissertation is organised and the concluding Section summarises this Chapter.

1.2 Institutional background

The Web log survey was conducted at a UK (English) college of higher education which awards its own degrees.[1] During the period of study this institution had about 8,000 students of whom about 5,500 were full-time undergraduates. The College profile is given as:

> The College has three faculties:
>
> Arts and Education
>
> Business and Social Studies
>
> Environment and Leisure.
>
> The major full-time undergraduate college programme is the modular degree scheme offering an extensive programme of subject choices. The largest areas of provision are in Business and Management Studies and Education.
>
> (Higher Education Funding Council for England, 1997, p. 204)

The overall gender ratio for all undergraduates at the institution is about 57% women which is slightly more than the corresponding UK figure (54%). During the academic years 1997/1998 and 1998/1999 there were 2,103 and 2,345 first year full-time undergraduates respectively. In 1997/1998 the age distribution of these students was

[1] The institution has subsequently been awarded University status.

68% aged under 21 years. This proportion increased to 73% for the 1998/1999 cohort. The corresponding UK figure is 69% (Higher Education Statistics Agency, 1999, 2000).

Hence it appears that the institution's undergraduate population may have a small female bias but that while the two cohort age distributions fluctuate about the UK average, taken together they are typical.

The 1,050 full-time undergraduates whose Web information seeking activity is the empirical basis of this investigation are drawn from the 4,448 full-time undergraduates in the 1997/1998 and 1998/1999 cohorts. This 24% sample is *all* of the these 4,448 students who used the Web as provided by the institution to seek Web information (that is not including email, OPAC, chat-room or internal academic support information) on at least two days during each of the two study-years. Throughout the dissertation they are referred to as *student-users*. The student-user gender composition is discussed in Chapter three. The Web access facilities provided are by way of computers situated throughout the institution's 'learning centre'. During the period of the study this was generally open seven days a week from 9am to 10pm or 5pm at weekends.

It was anticipated that more of the institution's learning material would be delivered electronically during study-year two than during study-year one but that this increase would not be uniform across all subject modules. It would therefore distort the investigation's findings. The *conditioning* process (see page 59) which discounts students' Web requests to the institution's own Web based resources controls for this distortion. However it is possible that an increase in learning material provision is a factor in stimulating an increased use (in terms of session rate, see page 64) of the publicly accessible Web.

Appendix A gives extracts from the *Students in higher education institutions* (Higher Education Statistics Agency, 1999, 2000, 2001) which shows an analysis of the areas of study of the institution's full-time undergraduates together with the available age information. It is not possible to determine the subject mix within particular cohorts. However during the 1998/1999 and 1999/2000 study-years the overall subject mix shows the preponderance of 'business & administrative studies' which accounts for over 20% of all full-time undergraduates during each academic year. It will be seen that as far as the 1,050 student-users are concerned, once use of any internally provided academic support information has been discounted, no 'academic' type Website appears in the 'top-twenty' popular Websites during either study-year. This is consistent with the feeling gained (anecdotally) while working with the Web logs that the observed information seeking was not predicated by the educational context of the institution but may have been equally observed within say, a 'cyber-café'. This, together with the meta-analytic form of analysis which is not driven

by Website content, suggests that students' subject mix is not an important factor when considering how they locate Web information.

1.3 Dissertation organisation

The dissertation is organised conventionally to the extent that it is approximately; introduction, literature review, methodology, findings, conclusion, reference list, and appendices. The substantive content of every Chapter is also prefaced by an introduction and concludes with a summary. Chapters one, two and three correspond to the conventional arrangement.

The literature review in Chapter two considers issues in *real world* Web research and research into the problems of analysing Web logs, in particular the difficulties of distinguishing users and sessions when carrying out *server-side* based research (which is generally more interested in the characteristics of the requests to a particular *Web server* and is not informed of what users are doing at other Web servers). Issues associated with caching are also discussed. Eight actual or surrogate *client-side* based Web research investigations which report what users do are reviewed in detail. Client-side investigations can be informed of all the Web servers that a user visits but not all the users who visit a Web server. Of the larger scale investigations user identification is reliable only in the study by Cothey (2002).

Chapter three sets out the research design, data collection, analysis and inference procedures. These are based on *Web log analysis* which resembles more conventional TLA but is in respect of a Web log. All the *cleaning* of the Web log file, record categorizing and frequency counting was achieved using computer programs which manipulate the log file written by the author. Analysis and interpretation of the data go hand-in-hand especially because of the exploratory nature of the study. Examples using the empirical data are therefore used to illustrate analytic techniques. A feature of the Web log analysis (which is also reported in the literature) is the strongly skewed distribution of some of the derived metrics.

Chapters four, five and six each report findings. These are of three types. Chapter four describes the findings of the repeat study and how the sample of student-users locate Web information is considered separately for each of the two study-years. Chapter five focuses on how student-users use Web 'search-engines'. In particular this includes an analysis of all the *queries* to two particular 'search-engines' from those student-users who used the 'search-engines' concerned.

Chapter six describes the findings of the investigation which are designed to examine the change by individual student-users in how they locate Web information over the

survey period. This is done by comparing what each did during study-year two with during study-year one. The design of this part of the investigation is called *longitudinal-developmental* (Nesselroade & Baltes, 1979).

All three of the Chapters relating to findings also discuss the rationale and construction of the characterization metrics which are used. These metrics inform the development of a narrative description which differentiates different groups of student-users by how they locate Web information. The narrative description also makes use of the idea of a *personal information infrastructure* as described by Marchionini (1995) when referring to student-user's *personal Web information infrastructures.*

The final Chapter summarizes the findings of the previous three Chapters and concludes how it is that the participating student-users locate Web information. Chapter seven includes also an identification of the limitations of the study and some suggestions for future work which arise from the thesis.

1.4 Summary

The Web is becoming pervasive but there is little scholarly understanding about how users use it. This knowledge vacuum is filled by unsubstantiated myth, anecdote and 'factoid'. Research of the Web is also faced with difficulties of terminology. The language and descriptions which accurately describe what the Web is, how it works and how it is used are the components of the various infrastructure programs which control its operation. These are generally inaccessible to all but a few specialist cognoscienti. In consequence an approximate popular terminology is normally used but this is ambiguous and inconsistent. Therefore this investigation needs a more closely defined terminology. For example the meaning of *Websit*e which approximates to a uniform resource locator (url) is defined so as to correspond to the scrollable Web page which is displayed by the Web browser.

A particular ambiguity arises from 'use'. In all but a few studies of the Web, the users or persons who are using the Web are not individually distinguishable. Most often it is only the client browser software residing in a particular machine which can be distinguished. Hence while it is relatively straightforward to measure the *usage* of a Web server from its server logs, for example a frequency analysis of Web page requests, it is usually not possible to say anything (either reliable or valid) about *use* or what it is that a user is doing. Without some purposeful intervention one cannot even count the number of users. Throughout the dissertation the meaning of 'use' and 'user' entail the notion of a distinguishable individual person but 'usage' is apersonal.

'How do students who use the Web locate information resources?' is taken to mean, what do students *do* when they are obtaining Web information? This question is operationalised in terms of students' Web information seeking actions or requests to a Website to provide a Web page for display by the Web browser. The action of a request is referred to as a *click*. Requesting Web information is often called 'visiting' a Website so that the investigation is a study of the frequency patterns of student-users' clicks as they visit and revisit Websites.

The empirical data comes from a UK higher education institution. The participation of this institution in the investigation has its origin in a previous study (Cothey, 1998) in which a sample of data is analysed which is similar to a snapshot of the longitudinal empirical data that is analysed in the present study. Prior to and during part of these studies I was employed by the institution and hence the study can be said to be opportunistic. During the period of the study the institution provided unrestricted access to the Web for its students from computers located throughout its learning centres. These computers were normally available between 9am and 10pm. The study is based on an analysis of the transaction log of Web information seeking activity (or Web log) from full-time undergraduate students. These students are referred to as *student-users* and each day's worth of Web information seeking by a student-user is a *session*.

The Web log is analysed both *by-user* and *by-session*. The by-user analyses consider separately the Web information seeking of each student-user. This is made possible by each student-user being anonymously identified in the Web log. The by-session analyses similarly consider separately each session. A particular form of analysis is the longitudinal-developmental analysis which considers change over time in Web information seeking by-user.

There are two confounding factors which may affect the investigation. The first is that the structure of the Web itself is subject to continuous change and secondly individuals may change as a result of becoming more proficient. Hence both structural and individual change may affect Web information seeking activity and are recognized as potential explanations for any longitudinal changes that are identified by the analyses.

The thesis that is presented here is that student-users develop a personal Web information environment which over time becomes more individually distinctive. As student-users' become more proficient then to an increasing extent their Web information seeking is located within a personal Web information environment and their 'searching' of the Web diminishes. This information behaviour is interpreted as student-users developing *personal Web information infrastructures* which notion is based on Marchionini's *personal information infrastructure* (Marchionini, 1995).

2

Previous research into how users locate Web information

2.1 Introduction

Web information behaviour is a special case of users' information behaviour more generally and Wang's (2001) review of *Methodologies and methods for user behavioral research* is complemented by Hsieh-Yee's (2001) review of *Research on Web search behavior* and *Measuring the Internet* (Molyneux & Williams, 2001).

Users' information behaviour embraces information seeking/searching which comprise the active components of behaviour (for example, Wilson, 1999) even though there is a lack of generally agreed precise terminology. *Searching* as an information behaviour is often reserved to mean just submitting information retrieval (IR) type queries to an IR system (for example Baeza-Yates & Ribeiro-Neto, 1999, p. 20) or to a so called Web search-engine (for example Jansen & Pooch, 2001). *Browsing* is used by Baeza-Yates & Ribeiro-Neto in polarised contrast to searching however Marchionini (1995) suggests that there is a more continuous spectrum of information seeking activity. His thesis is that browsing is a more natural mode of information seeking and that search (equals IR) systems should support more of this type of information seeking interaction (in *addition* to searching). Information seeking using IR search-queries is also described as being more *analytic* (Bilal, 2000, 2001; Marchionini, 1995; Qiu, 1993; Schater, Chung & Dorr, 1998) or specific (Chen, Wang, Proctor & Salvendy, 1997) compared to browsing. The use of search-queries is also associated with being *systematic* (Bilal & Kirb, 2002).

Empirical research in information seeking/searching can be categorised as being either experimental (laboratory) or naturalistic (Hsieh-Yee, 2001; Wang, 2001) which is sometimes referred to as "real world" (Lesk, 1998) or "real life" (Spink, Wilson, Ellis & Ford, 1998). Investigations generally apply a theoretical model (for example,

Marchionini, 1995) of information behaviour whereby information seeking depends on the interaction of several factors. Particular information factors are the;

task whether this is open-ended or is goal directed (Chen *et al.*, 1997; Kim, 2001), that is, has a specific correct answer, (goal directed tasks are expected to result in more analytic information seeking),

context or the system, the knowledge domain, and the circumstances which frame the information problem, and

individual differences especially information seeking experience.

The review *Research on Web search behaviour* does not restrict itself to Web seeking/searching per se; it reviews Web information behaviour generally and concludes that:

> Information seeking on the Web is a complex phenomenon ... Research conducted between 1995 and 2000 shows that researchers drew on earlier research on online search [sic] behavior and the theories and findings from related disciplines to investigate this phenomenon. Most research on children's search behavior described their interaction with the Web. Research on adult searchers focussed on describing search patterns, and many studies investigated effects of several factors on search behavior, including information presentation, type of search task Web experience, cognitive abilities, and affective states.
>
> (Hsieh-Yee, 2001, pp. 181–182)

The choice of factors influencing Web information behaviour noted by Hsieh-Yee reflects the continued use of earlier theoretical models although Jansen, Spink & Saracevic conclude that:

> ... Web search users seem to differ significantly from users of traditional IR systems, ... [which] points to the need for further and in-depth study of Web users.
>
> (Jansen *et al.*, 2000d, p. 226)

This suggests that care needs to be taken and previous findings may need to be revalidated (Chen & Cooper, 2001). Ford, Miller & Moss are even more emphatic as regards the distinctiveness of Web information behaviour and say of Web information seeking that it is:

> ... important to conduct information retrieval (IR) research within a Web-based context, as opposed to extrapolating from studies in more traditional information contexts, because the users of IR tools on the Web are very different from users of traditional retrieval tools.
>
> (Ford *et al.*, 2001, p. 1051)

In consequence of these concerns *real world* Web information seeking research is undertaken which avoids using an "imposed query" (Gross, 1995, 1998, 1999, 2001). In real world research users respond to their own information problems rather than Web information seeking in respect of an experimental task constructed by the investigator (compare for example, Carroll, 1999; Chen *et al.*, 1997; Schater *et al.*, 1998). Fidel, Davies, Douglass, Holder, Hopkins, Kushner, Miyagishima & Toney suggest that real world research, which is characterised by them as "analyzing users' seeking and searching behavior as it occurs in actual situations" (1998, p. 36) provides a better basis from which to study information behaviour.

The contrasting experimental procedure for studying information seeking has its origins in the work of Cove & Walsh (1987; 1988) who evaluated the information seeking affordances (Norman, 1983) of a system by characterizing how it was used in respect of information tasks specially constructed to prompt exploitation of these affordances. They concluded that a browsing affordance for online information seeking could be distinguished which was associated with an open-ended or non-specific (as opposed to a closed or goal-directed) information seeking task. It is generally considered that there is a causal relationship between task type (open, non-specific) and information seeking activity.

The *system* context of real world Web information seeking is the accessible Web[1] or some subsystem such as a particular *search-engine*[2] or Web information location service. Other real world contexts include a variety of knowledge domains and the possibility of *successive* information seeking in multiple sessions (Spink, 1996) whereby users develop their information problem and refine their seeking activity.

Real world Web research can be categorised as either *client*-side or *server*-side based (Molyneux & Williams, 2001). Client here refers to the Web browser software which is instructed by the user to request that files from a Web server be sent to the client's Internet Protocol (IP) address (for example, Nowick, 2001). Each server is aware of all the IP addresses from which it has received requests but is not generally informed of the other Web servers from which any client requests files. On the other hand a

[1] Defining the Web is problematic and exactly what is accessible will vary from between investigations.

[2] The precise definition of a 'search-engine' is problematic and is considered in more detail later in Chapter five.

client-side perspective of Web information seeking facilitates an awareness of each Web server from which files are requested (but not of all the clients requesting files from a particular server).

The seminal investigation by Catledge & Pitkow (1995), which is cited in all three reviews above, is a client-side transaction log analysis (TLA) of user's Web information seeking. It therefore responds to all of the three interleaving themes of this dissertation, namely real world information behaviour, TLA, and Web client side. This accords with the opinion of Fidel *et al.* (1998) that investigators should concern themselves with users' Web information seeking as it occurs in actual situations.

Five distinct schools of research into how users locate Web information emerge. The work at Xerox PARC[3] is the most prominent (Chalmers, 2000, 11 November) and includes Card, Huberman, Pirolli and Pitkow in the development of an ecological information foraging (Bell, 1990; Pirolli & Card, 1999) explanation of Web information seeking.

The Boston University OCEANS[4] group (for example, Barford, Bestavros, Bradley & Crovella, 1999), although apparently no longer active, developed mathematical models to interpret user's Website *vocabulary* growth[5] or *repertoire*[6] growth which they derive from *trace analyses* based on monitoring network traffic. The OCEANS group's work is focused on the technical effect which a user's information seeking has on the Web's infrastructure rather than on a user perspective of how a user locates Web information and therefore their contribution is principally methodological.

Investigations associated with Greenberg (for example, Cockburn & Jones, 1996; Tauscher, 1996) adopt an human-computer interaction (HCI) perspective. Greenberg (1993) is concerned with systems which he describes as *recurrent* and focuses on *history mechanisms* and users' re-use of interaction activity.

Fourthly there is the HomeNet project (Kraut, 1996) which is based on a panel of residential users recruited in order to support a longitudinal study of Internet usage.

The last school of research into how users locate Web information derives mainly from library and information science. It is represented by investigations into the use of search-engines, for example the Excite project (Jansen & Spink, 2000) and the Sheffield Web search strategy project (Ford, Wilson, Foster, Ellis & Spink, 2000), and by ad hoc studies such as school-student's use of the Web (Fidel *et al.*, 1998;

[3] Palo Alta Research Centre

[4] Object Caching Environment for Applications and Network Services

[5] Thomas (1998) uses the term *vocabulary* extensively to describe different system commands in his long term study of human-computer interaction.

[6] A user's Website repertoire is the cardinality of the user's Website vocabulary.

Schater *et al.*, 1998) or the integration of Web based information in decision making (Choo, Detlor & Turnbull, 1998).

The three Sections which follow consider research organised by the major themes of this dissertation, namely,

Web information seeking/searching particularly in the real world context but also including a discussion of information task and novice information seeking. Generally, reference to information seeking is intended to encompass information searching.

Web log analysis which examines TLA when applied to analyse Web information behaviour, and

Web client-side investigations which is a critique of actual and surrogate client-side information seeking studies. These fall into three methodological categories depending on the study's design,

snapshot based studies which may involve from a few hours to many day's worth of data but where the analysis amalgamates the data and does not consider information seeking phenomena which evolve that is are connected in some way to the previous information seeking of the individual.

extended studies collect data in order to investigate evolving information seeking such as successive information seeking. A particular extended phenomenon is vocabulary, for example, a user's Website vocabulary.

longitudinal investigations which compare information seeking phenomena at different times. Extended designs will not be longitudinal unless the (extended) phenomena are analysed with respect to their change over time. Similarly longitudinal designs will not be extended unless the phenomena studied are extended. *Longitudinal-developmental* (Nesselroade & Baltes, 1979) studies refer to longitudinal change in the behaviour of individuals as opposed to, for example, system usage.

Longitudinal investigations are not well represented in the literature (Yuan, 1997) and in particular only Cothey (2002) reports an extended longitudinal-developmental investigation of Web information seeking.

This Chapter also includes a brief discussion about the ethical practice of Web information seeking research and concludes with a summary of the previous research.

2.2 Web information seeking/searching

In this dissertation there is an emphasis on a real world context in which information seeking comprises users' (successive) information seeking responses to their self-generated information problems. This Section shows initially how researchers have responded to the demands of real world Web investigations. This is followed by a discussion of task and individual differences within the context of Web information seeking.

2.2.1 Real World Web information seeking

Investigations of real world Web information seeking require a study of real users resolving their own information problems in the Web context of their choosing. Our understanding so far is based mainly on the use of search-engines and imposed information tasks in an experimental context which relies on possibly atypical users (for example in order to investigate the effect of age difference). As Ford, Miller & Moss say when commenting on the difficulties of interpreting their experimental findings:

> ... clearly much more research is needed – in particular:
>
> 1. in more naturalistic, less constrained conditions;
> 2. using more complex and meaningful measures of relevance;
> 3. across a range of different types of search task;
> 4. in relation to a range of different populations of Internet users;
> 5. taking into account information seeking strategies as well as results;
> 6. making use of a range of different search engines.
>
> (Ford *et al.*, 2001, p. 1,063)

Given the theoretical model of task dependency and individual difference then real world information research is validated by scale. That is, it is assumed that a sufficiently large sample of information problems across a large enough user population will generate a validly representative sample of tasks of varying specificity undertaken by averagely typical users. Therefore real world information seeking corresponds to real world task specificity and users. Moukdad & Large (2001, p. 350) describes this as providing "more heterogeneous and therefore more representative samples" of information seeking activity. This contrasts with the experimental approach where the task is controlled and the number of participants is often quite small.

The experimental approach for investigating information seeking can be criticised on two main counts. Firstly it focuses on tasks within self-contained sessions of activity. This is unrealistic as Lin & Belkin (2000) say when they introduce the need to consider multiple episodes of information seeking:

> A big assumption about Information Retrieval (IR) systems is that information seeking sessions are discrete. This underlying assumption implies (a) that an information seeker must resolve his/her information problem in a single episode and (b) that information seeking activities in different episodes are unrelated. The first implication is problematic because in real life there are many factors that can keep information seekers from completing their tasks with in a single episode, ... [and the second is] problematic because the information problems and knowledge states of a person are temporally related, given that life is active, analog and accumulative.
>
> (Lin & Belkin, 2000, p. 133)

They argue that information seeking behaviour should be considered as being "problem centered rather than session centered" (Lin & Belkin, 2000, p. 133).

Spink and Spink, Griesdorf & Bateman also criticise research based on a single session approach and say that:

> Recent research exploring human information-seeking behavior suggests that the single search session model of end-user behavior has limitations as users move through a series of stages ...
>
> (Spink, 1996, p. 604)

and,

> ... successive searches are a fundamental aspect of user's behavior when seeking information related to an information problem.
>
> (Spink *et al.*, 1999, p. 479)

Secondly, the experimental approach uses imposed tasks so that in respect of the information problem being investigated:

> The information need or question is not his or her own in the sense that it was generated in his or her own mind or out of the context of his or her own personal life.
>
> (Gross, 1995, p. 236)

In addition the imposed tasks used in the experimental approach are frequently of different specificity, for example, either goal-directed or open *as judged by the investigator.* There are thus two further underlying implications when applying this experimental model; (a) that imposed and self-generated information seeking are the same, and (b) that the investigator and the information seeker share judgements regarding task specificity. Both of these are problematic.

The real world study of information seeking is in its infancy and, driven by expediency, the study of Web search-queries which can be based on analysing server-side query-logs has led the field.

The largest project (Jansen & Spink, 2000) to investigate real world Web information seeking was initiated as a consequence of the shortage of research highlighted at the 1997 conference "Real life information retrieval: commercial search engines" (Lesk, 1998). Jansen, Spink & Saracevic[7] accepted the offer by the Excite Web search service to analyse a query transaction log file. These analyses[8] form a "a major and ongoing study of user's searching behavior on the Web" (Jansen *et al.*, 2000d, p. 208). However, since the log files are server-side based and relate only to search-query submissions, the project's scope is to investigate only a particular aspect of Web information seeking, that is the submission of IR type search-queries. Hence the problem solving context of the individual has moved beyond using the Web in a broad sense, to using a subsystem of the Web. As yet we know nothing about whether or not this contextual shift is a typical problem solving behaviour. That is, do Web information seekers generally and uniformly use Web search-engine services?

In common with earlier investigations using TLA, for example (Borgman, 1986) the Excite investigation is based on *sessions, queries* and *terms.* These are defined as;

session an entire series of queries by a user,

query one or more search terms, and

term an unbroken string of characters (that is a sequence of characters not including a space).

Anonymized users are identified by the IP address of the client.[9] 51,473 queries are analysed and are classified as being unique, modified or identical (to another query

[7] Previously Jansen, Spink, Bateman & Saracevic (1998a,b); Jansen, Spink & Saracevic (1998c).

[8] For example Goodrum & Spink (2001); Jansen (2000a,b,c); Jansen, Goodrum & Spink (2000a); Jansen, Spink & Pfaff (2000b,c); Jansen, Spink & Saracevic (1999); Lau & Horvitz (1999); Ross & Wolfram (2000); Spink, Jansen & Ozmultu (2000); Spink, Jansen, Wolfram & Saracevic (2002); Spink, Wolfram, Jansen & Saracevic (2001); Spink & Xu (2000); Wolfram, Spink, Jansen & Saracevic (2001).

[9] This identification of the client is generally problematic but can be reliable during a short time interval when a user is continuously connected to the Internet.

coming from the same address) which contained on average 2.21 terms. However it is not clear how the queries, in particular the 43% of identical queries, have been processed. The Excite findings are discussed more fully in Chapter five where they are compared with an equivalent analysis conducted during this investigation.

Jansen & Pooch review the Excite study as well as a similar study of an AltaVista log (Silverstein, Marais, Henzinger & Moricz, 1999). They emphasise (2001, p. 237) three "major flaws" with the Excite investigation: (a) the query log is for only part of a single day and is therefore not longitudinal, (b) session identification is unreliable since, for example "computers located in public areas would have one identifier even though many users may have access to them" and, (c) null queries were not excluded so some analyses are in error. He & Göker (2000) are also concerned about the Excite (and AltaVista) session and user information:

> One major obstacle ... is the lack of relevant information about a user. For example, it is difficult to identify a user because a search engine usually does not have much information ... the IP address is not a reliable resource due to the use of proxy servers and dynamic IP allocation."
>
> (He & Göker, 2000, p. 7)

The AltaVista investigation (Silverstein *et al.*, 1999)[10] analysed almost a billion log entries collected over 43 days (but no attempt at any form of longitudinal analysis, for example query modification, is reported). The mean terms per query is 2.35 (sd = 1.74) and the maximum is 393 terms. The authors comment that 77% of sessions contain only one query and that their findings may be distorted by queries initiated by robots (Marshall & Roadknight, 1998) rather than human users.

As Hölscher & Strube comment, server based investigations have access to "impressively large data sets [which] give a detailed picture of how the average Web user approaches a search service, but they also have drawbacks: since the data are anonymous, we do not know anything about [the user]" (Hölscher & Strube, 2000, p. 338).

Unfortunately Hölscher & Strube are in error to suggest that we know anything about even the average Web *user*. As discussed in Chapter one, server-side studies in general can only provide information about average *usage*. But despite this limitation they do inform an understanding about *real* information seeking. Hölscher & Strube imply that client-side studies do not generally involve as many transactions or users but they do make it possible to consider individual differences as in their own study. Client-side studies of Web information seeking can be real world, but most are experimental. In addition to this difference in problem solving context,

[10] previously Silverstein, Henzinger, Marais & Moricz (1998).

Web information seeking experiments also make use of volunteer information seekers who are often information science students. It is not clear whether this self-selecting distinctive user group is representative of Web information seekers generally (Ford, Miller & Moss, 2002).

2.2.2 Task and individual differences

Chen *et al.* (1997), Schater *et al.* (1998) and Carroll (1999) report differences in Web information seeking related to task. Individual differences on how users locate Web information is investigated in several studies, (for example, as described by Chen, Czerwinski & Macredie, 2000). Novice-expert, gender, and age differences are all represented.

task

Task difference studies typically impose a pair of contrasting tasks which are constructed to be either "specific search"/"nonspecific browsing" (Chen *et al.*, 1997, p. 176) or "well-defined"/"ill-defined" (Schater *et al.*, 1998, p. 843). The studies are designed to measure the relative incidence of, for example analytic and browsing information seeking strategies (Schater *et al.*, 1998).

The methods and techniques used in task difference studies inform the design of the methods used for this investigation even though the particular information tasks of student-users are not studied.

All three of the studies considered here measured the total number of Website visits and Website revisitation (or clicks) which data Chen *et al.* collect by direct observation while Schater *et al.* use transaction logs (this procedure is not reported in detail). The study by Carroll analyses "meandering" and "hierarchical" (1999, pp. 217–218) Website visiting subsequent to using a search-engine service and is focussed on "success" (that is, locating the information expected by him as a response to the task imposed). He reports using video recording to capture the observational data (of eight participants) which is analyzed in respect of clicks but no detail is available, and concludes that "successful information retrieval on the [Web] is systematic" (1999, p. 220). Systematic in this context appears to mean not-meandering but the lack of Webhost/Website detail makes comparison with other studies impossible.

Chen *et al.* draws explicitly on Catledge & Pitkow (1995) and Marchionini (1995) when justifying path length as a measure since "the mean path length is supposed to measure the user's search strategy" (1997, p. 176). However the definition of the

start of each path is predicated on direct observational information and is therefore not compatible with the Catledge & Pitkow/Huberman, Pirolli, Pitkow & Lukose definition where the path length measures the number of different Websites visited within each new Webhost. Schater *et al.* rely just on the total number of clicks used to complete each task and do not report Webhost or Website information.

Chen *et al.* conclude from the observations of ten graduate engineering students that the "experiment confirms that different task types will lead to different user [information seeking] strategies" (1997, p. 177) or path lengths as defined by him. Schater *et al.* similarly discover from observing 32 school children that "Children employed significantly more analytic search strategies on the well defined finding task as opposed to the ill-defined searching task" (1998, p. 845) where by analytic search strategy they mean the use of a search-engine. Like the Carroll (1999) study, Schater *et al.* (1998) are mainly concerned with information task success or locating relevant information (as judged by a pair of adult raters) and the report also lacks detailed Webhost/Website information.

Despite there being insufficient detail to make arithmetic comparisons, all the studies illustrate using clicks as the basis of a measure of information seeking activity. They also all conform to a model which correlates inversely the specificity of the task and the number of clicks so that the less specific the task the more the number of clicks.

novice-expert

Defining and distinguishing expertise is problematic. In practise most researchers do not make the attempt and instead conflate the notion of novice-expert with that of tyro-experienced. For example, in *From novice to expert*, Mayer says:

> ... I use the term "novice" to refer to the least experienced group of users, [and] the term "expert to refer to the most experienced group of users
>
> (Mayer, 1991, p. 570).

Tabatabai & Luconi (1998) also explicitly conflate 'expert' and 'experienced' in respect of their three novices and three experts:

> ... we define expertise in terms of hours of Web access and not necessarily ... familiarity with the Web

so that:

> [The] novices' time spent on the Web was on average less than 3 hours a week whereas for experts it averaged 15 hours a week.
>
> (Tabatabai & Luconi, 1998, p. 390).

On the basis of observing for about an hour and a half how their six experimental participants used the Web to resolve an imposed task, Tabatabai & Luconi claim that the "experts had a more in-depth understanding of navigating strategies" (1998, p. 391) but no statistical analysis is reported.

When Lazonder, Biemans & Wopereis compared seventeen "novice" (less than ten hours experience) and eight "experienced" (more than 50 hours experience) users during three (imposed) tasks they found (2000, p. 576) that the experienced users "are more proficient at locating Web sites" but that "the performance of experienced and novice users was equivalent" on tasks that require users to find information on Websites.

Hölscher & Strube's client-side study of "experts and newbies" (Hoelscher & Strube (1999); Hölscher & Strube (2000)) is in two parts. The first part compares how twelve "established experts" (2000, p. 337) who use the Web daily in their workplace (and averaged 6.8 years Internet experience[11]) use a search-engine to answer imposed tasks with the average use of the search-engine. In the second part, 51[12] participants, half of whom are economics students, are classified as Web experts or not; "Web expertise was assessed by interview and pre-test, allowing us to clearly identify novices and advanced Web users" (2000, p. 343) but no detail is reported. From their analysis of how the participants answered imposed economics information seeking tasks it is concluded (but without reporting a statistical analysis) that Web media expertise and knowledge domain expertise can be separately identified and that "double novices" are especially disadvantaged when it comes to Web information seeking. For example, double novices were found to undertake more ineffective backtracking (2000, p. 343).

gender and age

Large & Beheshti (2000) and Large, Beheshti & Rahman (1999, 2002) use a case study design to investigate children's' successive Web information behaviour in respect of a classroom assignment. The investigation therefore escapes some of the methodological criticism (Meadow, 2000) relating to an imposed task and there being a correct answer which has been levelled at Lazonder *et al.* (2000). Large *et al.* (2002) refer to Schater *et al.* (1998) and analyse Web information seeking activity

[11] Christoph Hölscher: answer to question at WWW9, Amsterdam, 2000.
[12] Christoph Hölscher: answer to question at WWW9, Amsterdam, 2000.

based on analytic searching (equals IR type searching) and browsing (equals click frequency). For the 53 children studied, they report a gender difference in respect of both terms per query and click rate.

Mead, Spaulding, Sit, Meyer & Walker (1997) and Meyer, Sit, Spaulding, Mead & Walker (1997) conducted an experiment to compare how "eleven older community members aged 64 to 81 and fourteen university undergraduates" (Mead *et al.*, 1997, p. 153) undertook nine imposed tasks. They conclude that the older users were "inefficient" (in that they used more clicks than were necessary). This resembles the ineffective backtracking reported by Hölscher & Strube (2000).

The Sheffield Web search strategy project (Ford *et al.*, 2001, 2002) also identifies a gender effect. In their experimental study of 69 information science students' Web information seeking they find that, for example:

> The mean number of terms per query ... across both boolean and keyword queries were between 2.1885 and 2.3864 *except* in the case of keyword queries submitted by males, with a mean of 3.3473. As previously noted this difference was statistically significant.
>
> (Ford *et al.*, 2001, p. 1,060)

Despite the experimental studies which show some task and individual differences the effects are weak and confounded by other issues, for example a task difference revealed in school children's Web information seeking may not apply to adult users. However the research suggests a need to take account of potential task and individual difference effects.

Web information seeking research is in its infancy and a variety of research issues have been identified. As yet there is no established direction but there is a consensus of support for the position as summarised by Ford, Miller & Moss earlier (see page 16). This is that more naturalistic or real world research across a broader range of users is needed. Implicit within this position is the recognition of the developing importance of the Web as an emerging global information infrastructure (Borgman, 2000). This justifies the research effort to improve the accessibility of information whether through improved user training, better presentation or more effective search tools. Moukdad & Large (2001) recommend using TLA or *Web log analysis* to investigate real world Web information seeking. This is because TLA is able to support large scale naturalistic investigations. As mentioned previously, TLA is a major theme of this dissertation; previous research concerning TLA or Web log analysis is considered in the next Section.

2.3 Web log analysis

Transaction log analysis is discussed in a special issue of *Library hi-tech* (Peters, Kurth, Flaherty, Sandore & Kaske, 1993b) and more recently by Jones, Gatford, Do & Walker who say:

> Transaction logging has traditionally been accepted as a useful tool for monitoring the use of library and information retrieval systems.
>
> (Jones *et al.*, 1997, p. 35)

and further, that:

> A log consists of a sequence of messages written to a file or data base. Their selection and composition will depend on the functionality of the system being logged, the purpose for which the log is produced, the data-gathering mechanism used, and the intended method of analysis.
>
> (Jones *et al.*, 1997, p. 37)

Cooper discusses how this traditional technique might be developed to provide information about how users seek Web based information. His proposal draws draws attention to *cookie* files and server based logging both of which can make a contribution. However he points out that:

> ... what the Internet community calls logging bears no resemblance to the needs and requirements of the library and information science community. ... Nevertheless, these logs provide the basis for a veritable cottage industry of software developers who offer both public domain and commercial products to analyze the logs.
>
> (Cooper, 1998, pp. 911–912)

The extended version of the common logfile format, that is the extended set of information which is recorded by Web servers (and which are the logs referred to by Cooper, 1998) is specified by Pirolli & Pitkow as including:

- the time of request in seconds,
- the machine making the request as either the domain name or IP address,
- the name of the requested url as specified by the client,

- and a unique identifier issued by the server to each client (typically a cookie)

(Pirolli & Pitkow, 1999, p. 31)

As pointed out by Pitkow (1997, 1998) the interpretation of server log files is fraught with difficulty particularly when used to discover how users seek information, however the data are much used to aid Website design and to improve the commercial performance of Web based services (Burton & Walther, 2001; Novak & Hoffman, 1999; Nowick, 2001). Simplistic analyses are commonplace and are prone to error due to a lack of understanding of both the server logs and also the structure of the Internet.[13] For example, Pirolli & Pitkow point out that:

> Adding to the confusion, there is no standardized manner to determine if requests are made by autonomous agents (e.g. robots), semi-autonomous agents acting on behalf of users (e.g. copying a set of pages for off-line reading), or humans following hyperlinks in real time. Clearly, it is important to be able to identify these classes of requests to construct accurate models of surfing behaviors.
>
> (Pirolli & Pitkow, 1999, pp. 31–32)

The problems as regards reliable logging of user information seeking centre on identifying individual users (which is bound up with session demarcation) and *cache-busting*. The technique of analysis of a reliable transaction log presents a third problem area. Each of these areas is now discussed.

2.3.1 Identifying individual users

Both Pirolli & Pitkow (1999) and Silverstein *et al.* (1999)[14] discuss heuristics for identifying individual users in server side transaction log files. However their discussion is limited to identifying the client machine from which the transaction originates. Whose hands are on the keyboard cannot be detected generally, even in principle, by Web servers. This is not necessarily always the case with transaction logging since, for example, in networks where individual access is controlled by user-id/password protection then *in principle* the human user is identified by the user-id. The paradigm in server-side logging[15] is that it is sufficient to equate an individual user with each

[13] A frequent misunderstanding concerns the allocation of domain names which leads to the false presumption that country domain names correlate with geographic boundaries.

[14] previously Silverstein *et al.* (1998).

[15] This is in respect of publicly accessible search-engines. *Individual* commercial Web servers may employ user-id/password techniques.

client-session. Hence the usage of the so-called user is indicated by the set of transactions from a particular client during some self-contained time period or session. A subsequent session originating from the same client may be the same person or a different person but is generally deemed to be a different individual user. The server-side problem is therefore to reliably identify each client-session.

Silverstein *et al.* report that cookie information was available for over 96% of the search-engine queries which they studied and that the use of cookies is integral to their heuristic since in theory each user has a unique cookie. However:

> In actual use the situation is not so clear cut: different people using the same browser [or client] will share a cookie, and some users disallow cookies altogether. ... For those queries in which the user has disallowed cookies, we use the pair "domain IP/web browser used" as a substitute for the cookie.
>
> (Silverstein *et al.*, 1999, p. 7)

They admit that this substitution is a poor alternative particularly in respect of large Internet service providers such as America Online (AOL) where "tens of thousands of users can share a single IP" (1999, p. 7) but by comparing two data sets they claim that any bias is not material. (Since the comparison which they report is between the 96% of queries where a cookie is present and all 100% of queries then this is not surprising. This leaves open the question of whether the self-selecting group of users who have disallowed cookies also have different information seeking activity.) They continue that:

> ... a good sessioning algorithm [also] has to determine when a query starts a new information need. We use the heuristic that queries for a single information need come clustered in time, and then there is a gap before the user returns to the search engine. We used a cutoff of 5 minutes ... [but that since] in reality a user may try to fill two information needs in one sitting, our session-identification heuristic almost certainly underestimates the true number of sessions.
>
> (Silverstein *et al.*, 1999, p. 7)

Pirolli & Pitkow elaborate on the reliability of using the IP or domain name as a basis for identification since "within one session, a user may rotate between several proxies each with a different IP and domain name" (1999, p. 32) and refer to AOL as a case in point. They experiment with identification heuristics based on combinations of cookies, IP, domain name, timeout and "*host-munging*", a term which they coin to

describe transforming the host domain name for analysis by disregarding the host domain name prefix, and report that:

> ... the impact of the various [heuristic identification] strategies is quite dramatic and can sway the basic characterizations of session time and the number of clicks per session.
>
> (Pirolli & Pitkow, 1999, p. 35)

As an aside they say that their analysis of the server logs at `<xerox.com>` reveals that "weekend users [are] less likely to be cookie compliant than their weekday counterparts" (1999, p. 34) which reinforces the sceptical interpretation of the homogeneous view of users by Silverstein *et al.* (1999) noted above. Pirolli & Pitkow conclude that their analysis "presents empirical evidence that suggests the [different] methods used to identify users ... have [a] significant impact on basic characterizations of users' surfing behaviors" (1999, p. 37).

Identifying individual users or user-sessions is also bound up with the problem of determining session boundaries which is discussed below.

2.3.2 Cache-busting

The identification of so-called users and user sessions in server logfiles is predicated on the client being visible to the server. Not all transactions are seen by the server because (a) any Website request which the browser can satisfy from the local (client-side) cache does not reach the network, and (b) network requests may be routed through a proxy-cache which both hides the client from the server and can satisfy the client request without notifying the server. These circumstances may seriously mislead timeout based server heuristics for identifying individual users as well as generally attenuating server usage (which may be the design intent of a proxy cache (Smith, 1996)).

Cache-busting is a generic technique employed by Websites to oblige clients to always obtain a fresh copy of the url file whenever the browser issues a request to the Website. However a client hiding behind a cache is still invisible even though a cache-busting Website will be able to improve its estimate of the total number of requests made of it. Kelly has developed the idea and describes:

> ... a new technique for measuring Web client request patterns and [an analysis of] a large client trace collected using using the new method.
>
> (Kelly, 2002, p. 357)

His motivation is to understand and thereby improve network performance rather than to investigate user information seeking, however the technique he discusses could be applied to user information seeking research. Since his interest is the network he is as interested in client-side traces as server-side traces. Referring to Catledge & Pitkow (1995) and Cunha, Bestavros & Crovella (1995) he says that:

> In a few cases, researchers have instrumented browsers to collect Web client traces. In principle, such traces support arbitrarily realistic bottom-up explorations of cache hierarchies and shed light on user interactions invisible outside the client. ... In this [new] method, a "cache-busting proxy" intercepts requests from unmodified clients and labels all replies uncacheable, thereby disabling browser caches and allowing the proxy to log requests that would otherwise be served silently from browser caches.
>
> (Kelly, 2002, p. 357–358)

"True client traces" he continues (p. 358), that is "request streams not filtered by browser caches, are extremely rare ... and no true client traces have been collected since 1995". Hence he summarises the position as:

> ... the few existing Web client traces are several years old, reflect the requests of computer science students, and are small in comparison with server and proxy traces. By contrast, server and proxy traces are often large but typically omit much information in the original client request streams. e.g. references served from the browser caches."
>
> (Kelly, 2002, p. 359)

His criticism of the failings of previous research to inform an understanding of network performance is thus the same criticism that is applied here in respect of that research informing an understanding of user information seeking.

The massive (greater by "two orders of magnitude", p. 358) investigation undertaken by Kelly at WebTV Networks into cache performance found that:

> ... assuming *infinite* browser caches and perfect duplicate suppression, 73.4% of [the users'] requests would be served from browser caches. Of the remaining requests, 57.7% could be served from a sufficiently large shared proxy cache.
>
> (Kelly, 2002, p. 362)

which illustrates both the need to recognise the effect of caching, whether local browser caches or proxy caches when interpreting transaction logs, and indicates a projection for the overall scale of Website revisiting and between user Website commonality.

2.3.3 Web log analysis including session demarcation

Cooley, Mobasher & Srivastava (1999, p. 3) situate the problem of obtaining information from Web transaction logs within the domain of data mining and knowledge discovery (Trybula, 1998) and suggest that "mining for knowledge from log data has the potential of revealing information of great value". They identify (as has been noted previously) that "a raw Web server log does not reliably represent a user session file" and aim to address both "reliably identifying unique users and user sessions within a server log, and identifying semantically meaningful transactions within a user session".

Obtaining meaningful information is seen as a two stage procedure of data cleaning followed by data mining proper. They propose (1999, pp. 13–14) that data cleaning should eliminate requests for image type files except that " ... for a Web site that contains a graphical archive ... log entries of graphics file may very well represent explicit user actions, and should be retained for analysis". No advice is provided about how to do this. Data cleaning also provides user identification and session identification. "The goal of session identification" they say, "is to divide the [Website] accesses of each user into individual sessions" and that the simplest method is by using a timeout. Hence " ... if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session". They quote 30 minutes as justified by Catledge & Pitkow (1995) as being a popularly used timeout limit.

The data mining problem which they analyse is the problem of transaction identification about which they say:

> Each user session in a user session file can be thought of in two ways; either as a single transaction of many many [Website requests], or a set of many transactions each consisting of a single [Website request]. The goal of transaction identification is to create meaningful clusters of [requests] for each user. Therefore the task of identifying transactions is one of either *dividing* a large transaction into multiple smaller ones or *merging* small transactions into fewer ones.
>
> (Cooley *et al.*, 1999, p. 16)

Cooley, Mobasher & Srivastava construct three data mining algorithms and evaluate these by reference to both a real transaction log and a synthetic transaction log. Only one of the algorithms gave satisfactory results. In this context satisfactory means that the heuristic algorithm generated an acceptable statistical model of the usage of a particular Webhost. These analyses emphasise the impossibility of generally determining the information seeking of an *individual* user from server-side data; only average usage characteristics can be derived.

He & Göker (2000) and He, Göker & Harper (2002) are similarly concerned with the problem of analysing Web transaction logs and demarcating collections of transactions. However, while Cooley *et al.* (1999) are concerned with Web transaction logs generally, both He & Göker and He, Göker & Harper are interested only in search-engine query logs. In particular they distinguish their timeout period evaluations from timeouts based on Catledge & Pitkow (1995) since:

> ... the reasons behind choosing that particular amount of timeout are not clear, and the users' navigation patterns may have changed over the last six years. More importantly, Catledge and Pitkow's work is about users' navigation behaviour, and does not include activities of using Web search engines.
>
> (He & Göker, 2000, p. 3)

He & Göker (2000) found that optimal timeout values lay between ten and fifteen minutes. He *et al.* (2002) is a further analysis of the data used in He & Göker (2000) using artificial intelligence techniques in order to make use of the information of the content of the query or search pattern to improve the estimation of session boundary.

The motivation for analysing a Web transaction log varies and there are several communities of analysts who are interested in understanding how Websites within a Webhost are visited. Much of this interest is in relation to Web shopping (for example Fu, Sandhu & Shih, 1999), but there is also interest in Webhost performance (Schechter, Krishnan & Smith, 1998), Web information location/discovery (Cheung, Kao & Lee, 1998), and user behaviour (Chen & Cooper, 2001; Pitkow & Pirolli, 1999).

As yet there is no consensus as to the analytic procedures which should be used so the research is as much about presenting techniques as it is about revealing discoveries. Findings cannot be compared arithmetically because of the variety of technique and empirical or test data which is used.

Clustering techniques are frequently used in order to understand Webhost *usage*. Fu *et al.* (1999) invert this usual orientation and use a clustering technique to inform an understanding of *users*. They say:

> Our goal is to cluster Web users with similar access patterns. However, it is not easy to find many users who access common pages because of the diversity of Web users.
>
> (Fu *et al.*, 1999, p. 2)

Cheung *et al.* (1998) aim to characterize users on the basis of the similarity in the content of the Websites which they visit. Thus far the techniques which they report are experimental. The research is to develop heuristics which facilitate predicting Websites which are of interest to the user. This includes using a *term frequency*inverse document frequency* vector model (Salton, 1968; Salton & McGill, 1983) representation (tf*idf) of a consolidation of the collection of Websites visited for each user. Clustering bases include tf*idf topic vector clustering about a centroid where the topic vector is derived from a content analysis of the Website.

Su, Yang, Zhang, Xu & Hu (2001) examine the converse problem of clustering Websites by their usage based on the rationale that all the Websites visited by a user posses a degree of similarity. Their work is methodological in that it reports an experimental comparison of their technique with other clustering techniques.

The state of the art of Web log analysis is thus both primitive and sophisticated. It is primitive because the absence of reliable user and session information restricts what can be discovered about users while at the same time sophisticated techniques are used to glean information about overall usage characteristics. The research identifies pitfalls regarding the analysis of Web log data, in particular session demarcation, but there is no general agreement on either how to overcome these nor on methods for constructing meaningful interpretations of the data.

2.4 Web client-side investigations

The third major theme of this review of previous research is Web client-side and surrogate client-side investigations. These studies consider the Web requests made to all of the Web servers visited. Hence the goal of these investigations approximate to the objective of the research here which is to discover what it is that student-users actually do.

Eight real world Web client-side or surrogate client-side investigations have been identified. These are now discussed and are categorised by their design as being, snapshot, extended or longitudinal. Surrogate client-side studies use a data collection procedure which is not of itself client based but which intercepts (for example by using a proxy server) client-side activity.

2.4.1 snapshot studies

#1: Catledge & Pitkow Catledge & Pitkow (1995) identify the need to re-investigate experimental hypermedia information seeking in the context of real world Web information seeking. They therefore monitored over a three week period 107 staff and students at the Georgia Institute of Technology's College of Computing as they used a modified version of the Mosaic graphical Web browser. The modifications which they introduced captured client events and built a transaction log of users' Website requests. Web information seeking sessions were artificially constrained to last no more than 25.5 minutes and "users averaged 9.4 sessions each, or approximately one session every other day" (p. 1,068). However this session information is not used to support any of their conclusions.

In their analysis Catledge & Pitkow consider only requests to *external* Websites outside of the Georgia Institute of Technology and report that 1,222 different Webhosts were visited during their survey. They measured *path lengths* in respect of both within Webhost depth and micro navigational patterns. The Webhost depth is the path length of requests to Websites at a single Webhost (before visiting another Webhost) and is used in the development of the *Law of surfing* (Huberman *et al.*, 1998). Catledge & Pitkow report a mean of 10.31 (sd = 28.56) "successive document requests within a single [Webhost] across all users" (p. 1069). Also since there are 31,134 clicks then, on average there were 25.48 ($= \frac{31,134}{1,222}$) clicks per Webhost.

Computing their latter metric, the average path length per Webhost per visit, involves them employing pattern detection software to determine the frequency of occurrence of sequences of up to 50 Website requests. This form of analysis has not been repeated. Catledge & Pitkow conclude that Cove & Walsh's (1988) differential information seeking can be inferred also in respect of users' Web information seeking. This claim is founded on the observation that the frequency distribution of users' average path length per Webhost per visit is approximately linear (with average slope -0.24) but that some users "avoid the repetition of long invocation sequences" and therefore have slope $<$ -0.24 while other users "perform the same the same short navigational sequences relatively infrequently, but do perform long navigational sequences often" (p. 1,070) and have slope $>$ -0.24.

The importance of the study lies in it establishing the principle of measuring and comparing user's Web information seeking in respect of (cross-sectional) metrics such as path length. The authors' construction of session (by time limit) is also widely used.

Tauscher (1996) draws on Catledge & Pitkow, in particular by using the same modified Mosaic browser software, to investigate the design of browser history mechanisms and by carrying out a reanalysis of their data. Cockburn & McKenzie (2001);

McKenzie & Cockburn (2001) explicitly aim to revise Catledge & Pitkow in order to bring the findings up to date.

#2: Huberman, Pirolli, Pitkow & Lukose The Law of surfing study (1998) is a major snapshot designed study of Web information seeking. It is based on an analysis of data collected from an AOL proxy-server cache during a single day. The information seeking activity of 23,692 users provides a sample of 3,247,054 Website requests (which they also describe as clicks) from 1,090,168 Webhosts. Huberman *et al.* employ a user anonymizing technique and partially condition their data by excluding users' "requests for embedded media (such as images)" (1998, p. 96). Given how they analyse the data then the lack of Webhost conditioning is not important but it may need to be considered when comparisons are made with other studies.

They do not define explicitly all the terminology which they use and mix real client-side data with surrogate data when they amalgamate their data with Catledge & Pitkow so that, for example, the report "For the combined data, the mean number of clicks was 8.32 and the variance was 2.77" (1998, p. 96) is impossible to interpret. This is because their surrogate data is affected by local caching but the Catledge & Pitkow data is not.

Their principal metric is the path length or depth within each Webhost which was proposed by Catledge & Pitkow (1995). Since revisits satisfied by users' local caches will be invisible to the proxy server then the depth appears to be the number of *different* Websites visited at each successively different Webhost. The average depth value reported is 2.98 ($= \frac{3,247,054}{1,090,168}$) Websites per Webhost.

The Law of surfing describes the distribution of these depths. Huberman *et al.* (1998) show that the observed distribution agrees with a hypothetical distribution of surfing depths generated by a spreading activation (Pirolli & Card, 1999) algorithm. Thus the study provides an empirical foundation for the ecological foraging hypothesis posited by the Xerox PARC group.

#3: Kraut, Scherlis, Mukhopadhyay, Manning & Keisler The HomeNet project is described as being:

> ... an empirical field trial of residential Internet use whose goal is to increase our knowledge about the use and impact of residential electronic services. It uses longitudinal data collection techniques to study families' online behavior over time.
>
> (Kraut *et al.*, 1996, p. 55)

However the results reported relate only to a snapshot of 129 users of 54 weeks duration; Christ, Krishnan, Nagin, Kraut & Günther (2001) which is reviewed below, discusses an extended analysis of the HomeNet data.

The data collection procedure is described by them both as including "computer-generated use records of electronic traffic, newsgroups read and posted to, Web sites visited, and time on the Internet" (Kraut *et al.*, 1996, p. 56). Hence it is probable that client-side software monitoring is used to construct a transaction log. Interviews and pre/post trial questionnaires were also administered. It is reported that:

> Homenetters visited nearly 10,000 Web[hosts], but the modal Web[host] appealed to only one participant in the sample, implying that beyond a few highly popular services, people look for (and find) specific or niche, services matching their idiosyncratic interests. ... Of the 9,912 unique IP addresses visited, 55% were accessed by only a single individual and less than 2% were visited by 20% of the sample.
>
> (Kraut *et al.*, 1996, p. 58)

Nothing (other than the above) is reported concerning terminology or conditioning of the transaction log records. It thus appears that Webhosts were just converted to their IP address for analysis but since many Webhosts have multiple IP addresses (and some IP addresses support multiple Webhosts) this process is not reliable. The overall Zipfian (Zipf, 1972) characteristic of the distribution of Webhost popularity whereby in this case 55% of Webhosts are visited by a single individual only, is found generally. No evidence is presented to support the use of descriptions such as *appeal* or *look for* in the above quotations.

However, in spite of the shortcomings of the study as reported, the conclusion that "People gravitated toward services addressing their idiosyncratic interests" (1996, p. 63) and the notion of Web information seekers finding and occupying niches of Web information stimulates an interpretation of how users locate Web information.

2.4.2 extended studies

#4: Cunha, Bestavros & Crovella The OCEANS group are concerned with developing "efficient protocols to reduce Internet [the] traffic [following] the introduction of the World Wide Web and the explosion of network traffic attributed to it" (Cunha *et al.*, 1995, p. 1). But, as they say:

> Records of traffic to any particular server are readily available, as each server typically logs the requests it serves. However, server logs do not reflect the access patterns of individual uses.
>
> (Cunha *et al.*, 1995, p. 1)

The authors report that they therefore modified the Mosaic browser then in use in order to "acquire a complete picture of the reference behavior and timing of user accesses to the Web" (p. 2). They describe the record of urls visited as a *trace*. Comparative trace analysis profiles of two (out of 591) user's are presented which are based on repertoire trajectories; they say:

> Since we are interested in using user profiles and user past history ... we studied the rate of user access to new objects [equals Websites] in the Web. To do so, we plotted diagrams showing the patterns of individual users access to new and previously seen urls.
>
> (Cunha *et al.*, 1995, p. 8)

They identify that:

> If the user were continually accessing new urls (pure "surfing") the diagram [or repertoire growth curve] would have a slope of 1. If the user were continually accessing the same url, the diagram's slope would be 0.
>
> (Cunha *et al.*, 1995, pp. 8–9)

However the focus of their work is Internet performance so that these features are pursued as potential indicators for caching and not as indicators of an individuals Web information seeking behaviour. In the later work Cunha & Jaccoud (1997) propose that users' Web information seeking activity be distinguished by their evolving repertoire growth curves in order to adjust the operation of caching algorithms. The proposal uses a fractal random walk mathematical model (Thiébaut, 1989; Voldman, Mandelbrot, Hoevel, Knight & Rosenfeld, 1983) of the repertoire trajectory to identify perturbations, for example, when a user changes from mostly revisiting previously visited Websites to mostly visiting unvisited Websites.

Barford, Bestavros, Bradley & Crovella review the "caching properties of Web workloads" (1999, p. 15) and repeat the earlier trace data collection exercise during a seven week period during 1998. They used "non-caching http proxy software which recorded all request made by uninstrumented Netscape Navigator browsers" (1999, p. 16) which therefore provided surrogate client-side data (net of the local cache) and

collected traces from 306 users. In comparing the 1995 traces with those collected during 1998, it is reported that the average number of users visiting each Website had reduced from 2.57 to 1.27 and while the authors say the evidence is not conclusive it is suggested that, "relatively speaking, the most popular [Websites] are less popular in 1998 ... than in 1995. That is, [requests to Websites] in the 1998 dataset are spread more evenly among the set of [Websites]" (1999, p. 22).

#5: Tauscher Tauscher (1996) (also Tauscher & Greenberg, 1997a,b) draws on Catledge & Pitkow (1995) (but not Cunha *et al.*, 1995). In particular Tauscher uses the same modified Mosaic browser software, to investigate the design of browser history mechanisms. Tauscher also reanalysed Catledge & Pitkow's data in addition to monitoring the Website revisiting of 23 users (all computer science professionals) over six weeks. The users had to change from their usual browser software but this is not believed to have an effect. The advantage of the Mosaic instrumentation (although not reported) is that it captures all client Website access events (Barford *et al.*, 1999).

The phenomenon of particular interest is repertoire growth. The study is therefore extended because the analysis connects Web information seeking from an earlier session with Web information seeking during a later session. Since the data is analysed as a single continuous sequence of Website visits then session demarcation is not problematic. The data from five out of an original 28 users is excluded because they failed to record a minimum threshold of Websites visited (1996, p. 39).

Tauscher uses Greenberg's informal definition of recurrence and composition rate and in respect of the observed repertoire trajectories reports that:

> The most striking similarity across all 23 subjects is the regular increase in url vocabulary as portrayed by the approximately linear shape of the [trajectory] curve. The slope of the curve is roughly equal to each subject's composition rate.
>
> (Tauscher, 1996, pp. 57)

Like Barford *et al.* (1999) she identifies perturbations in the repertoire trajectory and correlates these with user activity such as, "*Revisits* to pages which produces an area of horizontal slope" and "*Authoring* [which] is usually apparent when a cluster of "*Reload* actions occur" (1996, p. 57–58). The inclusion of authoring suggests that the data has not been conditioned as regards the Web context, local or external. There is no argument presented to support a linear interpretation of the trajectory curve even though this is fundamental to her thesis.

#6: Cockburn & McKenzie McKenzie & Cockburn (2001) and Cockburn & McKenzie (2001) "aim to update and overcome some of the limitations of the prior empirical investigations into how the Web is used" (McKenzie & Cockburn, 2001, p. 1) by Catledge & Pitkow (1995) and Tauscher & Greenberg (1997a). They say of the earlier investigations that:

> ... there are [five] main reasons for suspecting that these findings may no longer reflect current use of the Web: the growth of the Web, the evolution of Web-navigation aids, the fact that the subjects were not using their "normal" browser, the relatively crude interface of the browser studied and duration of the evaluations.
>
> (Cockburn & McKenzie, 2001, p. 905)

The Cockburn & McKenzie study is based on an analysis of 119 days of daily client-side log files from seventeen computer science professionals which were obtained by extracting information from the institution's incremental backups of browser history files. The extraction procedure which they use to do this is improved for their second study although the base data are the same. The study was unobtrusive and essentially covert since their browser history file technique allows for retrospective data collection. When discussing this technique compared to one using instrumentation to log browser events they say:

> ... the technique we used to gather the data – file analysis from incremental backups – is different from that of prior studies, which used low-level logs of the actual user events (button clicks, etc.) executed at the browser. ... The primary strength of our technique ... lies in our ability to gather data about the user's browsing activities without changing, in any way, their browsing environment.
>
> (Cockburn & McKenzie, 2001, p. 920)

Their data analysis is based on clicks and Website repertoire growth. After four months the "mean per-subject final vocabulary size [or repertoire] is 1227 ($\sigma = 1086$), with a range from 74 to 4251" and "for each new [Website] added to the overall vocabulary, four [Websites] are revisited" (2001, p. 909).

Cockburn & McKenzie are almost unique in their providing details of their analysis methods and results saying:

> A C program was used to extract the data. To aid repeatability of the study, it is necessary to state four normalizations and assumptions made in the data analysis program.

(Cockburn & McKenzie, 2001, p. 907)

They also tabulate the full numerical findings for each of their 17 participants. This gives, for example, the regression slopes for the linear regression of visit count with repertoire.

A key finding is that "there was a surprising lack of overlap in the [Websites] visited by this fairly homogeneous community of users" (2001, p. 917), since:

> For each page in the total [Website repertoire] of 17,242 distinct [Websites] visited by the subjects, we counted how many subjects had visited it. Ninety-one per cent of the [Websites] had been visited by at most one of the subjects: that is, only 9.2% had been seen by more than one subject. No page had been visited by all the subjects, but one (the University's home page) had been visited by all but one subject. A total of 732 [Websites] had been visited by three or more subjects, and only 89 [Websites] were visited by eight or more subjects.
>
> (Cockburn & McKenzie, 2001, p. 917)

They thus conclude that Website revisitation is much more prevalent than had been supposed with about 81% of Websites visited by a user having been previously visited by that user and that there is a marked lack of commonality between the Websites that are visited by different users.

2.4.3 longitudinal studies

#7: Christ, Krishnan, Nagin, Kraut & Günther Christ *et al.* (2001) also (as well as McKenzie & Cockburn (2001)) presented an analysis of Web information seeking at the Hawaii international conference on system sciences. As mentioned above, Christ *et al.* use transaction log data collected by the HomeNet investigation to construct what they refer to as *developmental trajectories*:

> A developmental trajectory describes the developmental course of a behavior over time. Here we apply this method for the first time to the "development" of [Web] usage. We focus on the analysis of the number of *distinctive* Web sites accessed over time ...
>
> (Christ *et al.*, 2001, p. 2,795)

Unfortunately they do not define their terms but it appears that Web site means Webhost and distinctive means different (possibly different IP).

The authors observe that "in contrast to the exponential growth in Web[hosts] available ... there is actually a large decline in the average number of distinctive Web sites accessed" (2001, p. 2,796) but that different users may exhibit different kinds of access behaviour. Their analysis finds four groups of users which they label, non-users, moderate users, heavy users and very heavy users. 50% of the sample of 339 individuals are classed as non-users "who, but for a few visits to Web[hosts] immediately after the start, basically did not use the [Web] throughout the observation period" (2001, p. 2,797) of 144 weeks.

Christ *et al.* interpret their findings as meaning that users' Web information seeking becomes *saturated* over time so that each user reduces the number of different Webhosts which he visits each week. For example:

> The second group of individuals - moderate users - start [Web] usage at a higher level [than non-users] and follow a downward path in [Web] usage to a point of saturation between 3 and 4 distinctive Web[hosts] per week. This group is estimated to constitute 35.5% of the population. Among the population that actually uses the [Web] (non-users excluded) it accounts for 70.9% of the population.
>
> (Christ *et al.*, 2001, p. 2,797)

Both Christ *et al.* and Cockburn & McKenzie are considering Webhost repertoire but whereas Cockburn & McKenzie examine overall repertoire growth, Christ *et al.* are examining changes in *weekly* Webhost repertoire. Hence the analysis in this study is longitudinal in that the same phenomenon (in respect of each individual) is compared from one week to the next. However they do not connect together an individual's vocabulary from one week to the next so it is not known whether each week the same different Webhosts are being visited or not.

Christ *et al.* (2001) also report a gender difference but do not present any statistical analysis.

The study stands out because it demonstrates considerable heterogeneity of Web information seeking. Most users make very little use of the Web. Even when non-users are excluded, 80% ($= \frac{120}{165}$) of users are saturated at fewer than four Webhosts visited per week. But a few users, 9% ($= \frac{12}{165}$), "settle into a usage rate of about 50 sites [equals Webhosts] per week" (2001, p. 2,798). It also introduces individual Web information seeking *trajectories* as a useful analytical technique for studying how users locate Web information.

#8: Cothey The preliminary study by Cothey (2002) combines an analysis of the overall Website repertoire as undertaken by Cockburn & McKenzie (2001) with a longitudinal comparison which is based on a split half (King, 1991, p. 365) partition of each user's transaction log data. The data collection also resembles Cockburn & McKenzie in that it is likewise based on the browser history file.

The study is an unobtrusive client-side survey of the Web information seeking of 206 full-time students (98 men; 108 women) throughout their second complete academic year (about ten months) at a UK higher education institution. The study avoids one of the pitfalls in interpreting Web logs by defining each *session* as being all the activity from a user during a single day. There are 5,431 such daily sessions. Because the study uses the browser history file it also avoids all the pitfalls associated with local caching.

The study is unique since it is an extended longitudinal developmental investigation of Web information seeking. That is, an extended phenomenon of Web information seeking in respect of an individual is compared over time. In order to do this Cothey uses conditional regression and refers to Plewis (1985) saying that "conditional regression offers a standard statistical technique for analyzing the phenomenon of change over time in an individual" (2002, p. 72). The rationale for the technique is that although a particular information seeking metric for a group of users may vary enormously (for example because of individual differences), for any particular user, the value of the metric in respect of some later time will depend mostly on its value at some earlier time. The statistical procedure examines how justified one is in using the earlier metric to predict the value of the later metric. In a situation of no longitudinal-developmental change then the value of the later metric would equal (on average) the value of the earlier metric.

Cothey (2002) also uses a vector model based technique to examine "the degree to which a user chooses to access the more popular Webhosts, that is Webhosts accessed by most users" (2002, p. 74). In this study users are reliably identified so that it is possible to determine Webhost popularity in this way rather than using some usage related measure. Different so-called popularity measures are discussed in Cothey (1998).

The study found that as users became more experienced so they became more *passive* (they relied less on formal or search-engine querying) and more *eclectic* (selecting just a few Websites from a larger collection overall) in their information seeking. He concludes that:

> It appears that each user may inhabit an individual niche of Webhosts that become more distinctive as the user becomes more experienced. ...

> This suggests that it is misleading to assume that there is a uniform large scale homogeneity in Web users' information searching behaviour. The empirical evidence points towards users becoming increasingly distinctive regarding the Websites that they use.
>
> (Cothey, 2002, p. 77)

Cothey (2002), like Cockburn & McKenzie (2001), is a reliable study of users' uninhibited information seeking. Cothey (2002) compared with Cockburn & McKenzie (2001) is more extensive (206 users compared with 17), extends over a longer period (ten months compared to four) and is longitudinal but is limited in that it considers users' information seeking at the level of only Webhosts rather than Websites. The host-munging procedure which is described (Cothey, 2002) in order to clean the data is more extensive than Pirolli & Pitkow (1999) and more reliable since it equates the domain name server (DNS) *official name* (Mockapetris, 1987) of Webhosts. Cothey (2002) also (independently) uses the corrected technique for reading the browser history file as reported by Cockburn & McKenzie (2001).

As noted previously, it is difficult to compare arithmetically the findings from the eight studies because of the variety of techniques, metrics, motivation and reliability. However Cothey (2002) draws attention to the similarity of his overall conclusion with Christ *et al.* (2001) and Cockburn & McKenzie (2001). Christ *et al.* (2001) appear to show that week to week users typically each tend to visit fewer different Webhosts and Cockburn & McKenzie (2001) are surprised by the lack of overlap or commonality in the Websites visited by their homogeneous user community. In addition Kraut *et al.* (1996) suggest that users gravitate to Websites which support their own idiosyncratic interests. Taken together, these conclusions support the notion that users develop personal Web information environments or become characterised by information seeking involving an individually distinctive collection of Websites.

The next Section compares briefly the ethical practice of how the studies were undertaken.

2.5 The ethics of Web information seeking research

The ethical practice of the eight investigations discussed varies. Not all address the topic explicitly which appears to be related to how remote the researcher is to the person from whom the information seeking data originates. Also, as might be expected, whether the ethical practice is explicitly reported appears to depend on the ethical mores of the primary audience for the research. Hence practice is always explicit in respect of work reported in psychology but it is infrequently reported in

computer science. Jansen & Pooch make no reference to ethics at all in their proposal which sets out to "develop a methodological framework ... for future Web searching studies" (2001, p. 235). It also seems that reporting practice depends on whether or not explicit principles of data protection apply, so for example, European type practice varies from practice in the US.

Two models emerge, the *consent* model and the *permission* model. Experimental investigations invariably apply the consent model. This may take the form of participants volunteering to be observed, or being requested to consent to being observed. In these instances consent is prospective which in an experimental setting is not problematic but for real world investigations prospective consent may influence the outcome. Hence Cockburn & McKenzie use retrospective consent and cite Mayo (1933) commenting that:

> There were, therefore, no dangers of "Hawthorne Effect" modifications to subject behaviour due to their awareness that their actions were being logged.
>
> (Cockburn & McKenzie, 2001, p. 907)

The practice in investigations, both experimental and real world, where there are a small number of participants who are known to the investigator (for example because they are colleagues) and which apply the consent model contrasts with the (implied) practice of the larger scale more remote investigations. These latter investigations are also distinctive in that the researcher is not directly involved in capturing empirical data. For example the Excite studies are based on data which is made available to the investigators by the Excite service. The use of these data is therefore by permission of Excite and the question of whether of not an originating user has consented is not addressed. This is the case also with the AltaVista and AOL studies from the US and the Fireball study (Hölscher & Strube, 2000, p. 340) in Europe.

A tacit acknowledgment of the need to recognize potential ethical difficulties when undertaking this type of Web information seeking research is sometimes indicated by a reference to "anonymous" data, for example Huberman *et al.* (1998, p. 96) but this is not ubiquitous. In the US a practice designed prima facie to protect privacy appears sufficient and the literature is silent otherwise. In Europe the data protection principles provide more substantive guidance so that, for example, individuals should be generally protected against information revealed by analysis being used to support a particular decision concerning them. This appears not to be case in the US.

Other than the (possibly superficial) resolution of privacy concerns, large scale Web information seeking research is not so far ethically challenged. This is probably because the nature of the investigations is meta-analytic so that, for example, the focus

of attention is on *patterns* of Website visiting not on the Websites themselves. As a consequence the ethics of large scale investigations have not yet been explored and practice is dominated by a pragmatic approach founded on permission and privacy.

2.6 Summary and discussion

Research into users' Web information seeking is in its infancy and a coherent literature has yet to emerge. The current fragmentary nature of investigations is in part a consequence of the investigators' varied backgrounds which span fields as diverse as electrical engineering, computer science and psychology. They therefore have a variety of different research goals. There is a dearth of research aimed at discovering how users use the Web to locate information and such findings as there are, are often not central to the research undertaken. The majority of investigations have been engineering led where the research need is to discover how the Web's infrastructure or Web browser is used in order that these can be improved.

The fragmentary nature of existing research is exemplified also by the general lack of repeated results. This is manifest both by the research design which is rarely reported so as to facilitate replication, (Cockburn & McKenzie (2001) is an exception) and by the particular nature of the metrics used which tend not to be transferable. Many of the metrics which have been used are also criticised within the literature for lack of both reliability and validity, in particular in connection with the effect of the local browser cache.

There is a consensus among interested researchers that we are largely ignorant about what it is that Web users do and that more research should be undertaken. As yet the findings of Web information seeking research have not been related to a theory of users' information seeking which attempts to explain how users may change or develop their information seeking. This is because our ignorance is profound in that not only do we not know what Web users do, we lack both an empirical and theoretical frame of reference within which to construct our research. That is, we do not yet have a consistent set of definitions or metrics which can be used to characterize how users locate Web information and which would therefore support comparison, nor do we have any explanatory theory to critique empirical findings and direct further empirical investigations of Web information seeking.

The tentative empirical framework for Web information seeking which is emerging is based on users visiting (or requesting files from) Websites. This event is commonly referred to as a *click*. Clicking to a Website which has been previously visited gives rise to a Website revisit while the path length counts the number of clicks in some

sequence of Website visits. These ideas inform the design of the method which is used here and which makes use of closely defined metrics to characterize what student-users do.

It is reasonable for Web information seeking to lack any general theoretical support given the hybrid nature of information science (including for example both librarianship and information retrieval) which itself lacks any unifying theory (Rayward, 1996; Wersig, 1993). The field of information science is instead pragmatic and comprises localised theories and terminologies which may not be generally consistent (Cool & Spink, 2002) and may even conflict.

The information foraging theory advanced by Pirolli and his colleagues at Xerox PARC is only claimed to predict the observed Zipfian power law distributions of the empirical metrics. Foraging theory seeks to explain the behaviour characteristics of groups of users. Marchionini (1995) constructs a theory of "*personal information infrastructure*" (1995, p. 11) to frame his discussion of how individual users seek information in (traditional) electronic environments. He suggests that as users become more experienced and strengthen their personal information infrastructure so their information seeking strategies will become more *systematic.* Marchionini does not specify how systematic may be operationalised other than to suggest that users will rely less on browsing. The argument for greater systematicity is based on users developing their knowledge of how information in their domains of interest is organised.

The previous research into users' Web information seeking indicates that this is a complex phenomenon and that how Web users locate information may be significantly different from how users locate information in more traditional information environments. There is thus a need to develop an understanding of users' information seeking within the real world Web context. Marchionini's personal information infrastructure adapted to the Web context as a *personal Web information infrastructure* offers a basis for explaining or interpreting discoveries about how Web users locate information.

The literature also stresses that research into how Web users locate information should adopt a problem centered rather than a session centered approach so that users' extended information seeking which spans multiple sessions is considered. This can only be undertaken if users and their sessions can be reliably identified.

3

A method for discovering how student-users locate Web information

3.1 Introduction

It is not yet clear what is meant by *locating Web information.*

As discussed in Chapter two Web information behaviour is validly studied by investigators from many fields. Within the fields of library and information science Web information seeking/searching refers to the *actions* or observable components of Web information behaviour but *seeking* and *searching* lack precise general definition. Neither is what is meant by *information* precise so that, for example, it may restrictively identify only material which satisfies an immediate need, or it may inclusively embrace material which of itself fails to satisfy the need at hand but which leads the user to satisfying the need.

Addressing the semantic and philosophical issues which can be raised here is beyond the scope of this dissertation. As understood here an item of Web information is equivalent to a Web page[1] (of arbitrary length) which a user requests for display on his (or her) graphical Web browser. In this regard the method resembles that of Huberman *et al.* (1998) which is discussed in Chapter two. The *intent* of the user is not considered, therefore all Web information which is *requested* by a user is valued equally whether it be consumed as entertainment or education. This appears to contrast with Wilson, Ellis, Ford & Foster (2000) who understand *information seeking* as an active mode of information behaviour compared to "the passive reception of information as when a person watches television advertisements" (p. 1), but I assume that the watching of the television program is purposeful and is therefore included.

[1] By Web page I mean the collection of various computer files and data which are statically or dynamically assembled as a single viewable entity.

Like Wilson *et al.* Web advertisements which are not expressly requested are excluded. Marchionini (1995, p. 5) uses a definition which is similar; " ... *information seeking* [is] a process in which humans purposefully engage in order to change their state of knowledge".

Web information *seeking* and Web information *searching* are taken to be synonymous (although seeking is generally used) and denote user (Web information) actions. Hence Web information seeking is the collection of (Web) actions employed by the user to purposefully request Web information. The intent of the focus on Web information in this definition is to exclude for example the hand/eye movements studied by HCI practitioners. Web information actions are effectively either, (a) link-clicking which is the user clicking on a Web hypertext link or, (b) entering a uniform url or query term from the keyboard.

How do users locate Web information is thus given the meaning, what is it that users *do* when Web information seeking or what are the patterns of combinations of Web actions that users employ when requesting a Web page? In the context of this dissertation *users* is restricted to *student-users*, the population of users which is being investigated.

The focus on Web information in the context of Web information seeking also provides the rationale to exclude the use of the graphical Web browser to access chatrooms and email since neither is *Web information*. Operationally, Web information includes all other Internet information resources accessible via the graphical Web browsers provided by the institution. This is commonly referred to as being *the Web* but difficulties of terminology arise and qualifications such as the *publicly indexable*[2] Web, the *deep*[3] Web and the *dark*[4] Web are introduced (Bailey, Craswell & Hawking, 1999; Lawrence & Giles, 1998, 1999; Molloy, 1998). The Web referred to in this dissertation is greater than the publicly accessible Web since it includes resources made available through academic subscription. However, in practice it is seen that most activity takes place in the publicly accessible Web.

The method also includes a replication of the Excite search-engine studies, see Chapter two. This allows comparison of the *Web searching* (used here in a restrictive sense that is IR type searching) of a sample of student-users with the Excite Web searching findings.

Chapter three has five Sections which are;

[2] Accessible by search-engine crawler robots.

[3] Publicly accessible via Web *search forms* which information is therefore not indexed by search-engines.

[4] Intranets and other networked resources to which access is generally restricted.

Methodological overview which is an outline of the research design, data collection and analysis, and inference procedures that are used. This Section also includes a discussion of how the ethical issues arising from the research are resolved. The principal technique that is used is Web log analysis which is carried out twice, once on each of two transaction logs of approximately ten months duration. The results of each analysis is then compared including a *conditional analysis* in order to investigate *longitudinal-development.*

Web log analysis is TLA applied to Web logs. This Section describes the primary and secondary (meta) analysis of the Web log and how the reliability of the analysis is improved by *conditioning* or standardising the representation of *Websites* (which approximately equals urls).

Web vocabulary and trajectory analysis which examines models to describe how student-users develop their Web information seeking in respect of the number of different Websites which they visit.

Web session-conformance which discusses a technique to compare Web information seeking sessions, and

Longitudinal-developmental analysis which discusses in particular using conditional-regression to investigate how student-users change how they locate Web information.

The method also includes considering Website popularity and a sample of *search-queries.* For these the Web log is analysed more directly compared to the mostly meta-analytic approach used elsewhere within the method. These more direct techniques which are used are described and discussed in the context of the results of the popularity analysis in Chapter four and search-queries analysis in Chapter five.

The method is summarised and discussed in the concluding Section of this Chapter.

3.2 Methodological overview

This Section is an overview of the research method of this investigation in respect of;

1. research design,
2. data collection,
3. data analysis,

4. inference procedure, and

5. ethical issues.

The research problematic (Brown & Dowling, 1998) is Web information seeking activity, thus the investigation is a study of the observable (Web) actions of users which comprise their interaction with the Web information resource while they are searching/seeking information. Hence the research question, **how do student-users locate Web information** by which is meant, what are the patterns of combinations of Web actions that users employ when requesting a Web page is operationalised as, **what is student-users' Web information seeking activity**? Thus the question asks, what are the similarities, differences and changes in the patterns of clicks or visits to *Websites* by student-users?

As employed here, *to locate* and *to seek* Web information are equivalent in meaning and are used interchangeably. The investigation mainly concerns patterns of Web information seeking in a meta-analytic sense so that the analysis is an analysis of the analysis of urls and the navigational patterns of users between urls. An example of a meta-analysis is that undertaken in connection with the Law of surfing which:

> ... describe several strong regularities of Web user surfing patterns ... that [determine] the probability distribution of the depth - that is, the number of pages a user visits within a Web site.
>
> (Huberman *et al.*, 1998, p. 95)

The terminology commonly used to describe the Web is not sufficiently precise or consistent therefore some special terminology, such as *Website*, is introduced. A Website is approximately indicated by the url which is used to request the Web information. The procedure for determining the Website is defined below in the Section which describes data collection.

3.2.1 Research design

The research question focuses specifically on the user since the investigation is interested in what a user does, *not* in how the Web is used. Therefore any description should describe a user (Pirolli, 2000) and not a session. This focus is a key distinguishing feature of the investigation compared to how large-scale Web research is generally undertaken.

The design is thus naturalistic (rather than experimental) in order to discover real world information seeking. According to the taxonomy proposed by Tashakkori &

Teddlie the design is a mixed (or a combination of both qualitative and quantitative) methodology and is a methodological triangulation (Denzin, 1978) of (a) an "exploratory investigation, quantitative data and operations, statistical analysis and inference" design with (b) an "exploratory investigation, qualitative data and operations, statistical analysis and inference" design (Tashakkori & Teddlie, 1998, p. 145-146).

The design could also be described a longitudinal descriptive cohort study based on a manifest content analysis of the Web log (Holsti, 1969).

The exploratory or descriptive nature of the design is predicated by both the research question and the existing state of knowledge in that there are no a priori general hypotheses to confirm or reject. However the previous research suggests a qualitative framework for description which distinguishes both gender and usage frequency so that, for example, men and women student-users may seek Web information differently (although the parameters of these differences are unknown!).

Since in respect of this dissertation relevant Web information seeking parameters are not known in advance and therefore the homogeneity of users cannot be judged, it is essential that the investigation unobtrusively captures as fully as possible what individual student-users do. This is discussed below in the Section describing data collection. An unobtrusive approach is needed since their is some evidence that users modify how they use the Web when they know that they are being observed (Graham-Cumming, 1997, 1998).

The design is extended (or longitudinal) which helps to disambiguate *structural* influences, that is the effect on Web information seeking of the Web's evolving structure and information seeking affordances. For example, suppose a survey of users' Web information seeking activity concludes that this activity has certain characteristics and that another similar survey concludes that these characteristics are different, then does this indicate a change in Web structure or a change in Web information seeking activity (or both)? The longitudinal design of the investigation observes student-users' Web information seeking activity over two academic years. Although it is not possible to identify the effects of structural change with certainty, change effects which apply selectively to some groups of student-users but not to others are less likely to be structural in origin than change effects which apply to all the student-users. Different groups of student-users are differentiated on qualitative grounds of gender and also (using a qualitizing procedure) by how much they use the Web and how much they share the Websites which they visit with other student-users.

Students are also differentiated by their registration year cohort. The sample-frame for the investigation is all the full-time undergraduate students at the institution

from the 1997 and 1998 cohorts. The sample comprises all those who used the Web to seek Web information during two or more days within each of the two academic years 1998-1999 and 1999-2000. These periods are referred to as *study-year one* and *study-year two*. During study-year one students in the 1998 cohort are therefore in their first year at the institution and during this period they are here called *novice*. Novice as used here entails no inference regarding expertise, so that novice means not having experience, *not* not having expertise. Thus non-novices have experience but nothing is claimed about this increased experience being associated with increased expertise.

The investigation's primary data (described below) is the Web log of information seeking activity. This is mainly collections of Websites and is initially analysed to provide frequency data, for example the count of the number of different Websites which a student-user visits. This derived data which relates to each student-user provides data for the secondary or meta-analysis of the Web log which, for example uses statistical analysis to compare how many *different* Websites student-users each visit and how this changes over time.

As implied above the primary data includes student gender data and also search-engine queries. These data are linked by the individual (anonymized) student codes. Conclusions in the form of narrative profiles are formed by both statistical hypothesis testing (conjecture/refutation) and on the weight of triangulated evidence. This is discussed later.

3.2.2 Data collection

The data of the investigation consist of both the primary observational data and the derived secondary data which facilitates the meta-analytic aspects of the research design. Only the primary data collection is discussed here the goal of which is to unobtrusively survey over a two year period student-users' Web information seeking activity, that is, which Websites student-users visit and when.

As with all longitudinal investigation, attrition (Bijleveld, van der Kamp, Mooijaart, van der Kloot, van der Leeden & van der Burg, 1998) is a problem in that it cannot be known at the outset who from the commencing sample will still be present at the conclusion. Therefore the user sample-frame consists of every full-time student at the institution from the 1997 and 1998 cohorts. This large scale (4,448 students) and prolonged period dictates that an automatic procedure be found to survey or monitor for each individual all the urls which that student requests. Not every student in the sample-frame used the Web during the survey, those who did are referred to as

student-users. Only full-time undergraduate students are included in the sample-frame because it is considered that they represent a more homogeneous group, for example as regards age distribution and information task.

As discussed in Chapter two, server-side data collection procedures which are used to investigate usage generally fail to obtain information about a user's information seeking when Web requests are satisfied by the browser's local cache. In small scale client-side *machine* based studies this problem can be overcome by using additional browser monitoring instrumentation but no suitable instrumentation software exists for a large scale *user* based study.

However the institution's networked configuration of workstation terminals possesses three features which taken together almost completely fulfill the pre-requisites for the collecting the survey task. The institution's network infrastructure provides centralised application software supervision, user authentication and nightly data backup so that,

1. each student-user has an individual Web browser *global history* file which supports the graphical Web browser application's marking of Web hypertext links which have been visited. The global history file is a proprietary database which records user's cumulative visits to each url and hence the decumulation of a daily sequence of global history files reveals by subtraction the incremental addition made each day. This includes urls satisfied from files in the local browser cache.

2. network access is controlled by unique personal student-id and student-id/password combinations. Central data storage space which contains the browser global history file is allocated to each student-id and is not accessible by any other student-id.

3. nightly, out of hours, the centralised data storage space is copied to archive backup tapes so that generating a temporary additional copy of each student-users' history file is unobtrusive and not (technologically) onerous.

Decumulation of the nightly copies of each student-users' browser global history file (nearly) fulfils the goals for the data collection survey. The decumulation procedure is set out below together with a filtering procedure since this survey technique captures and records for each student-user all the urls requested both expressly and implicitly.

The goal of the survey is to record all of a user's Web information seeking activity but the data collection procedure fails to survey the time of *every* Web request because the global history file records the times only of the last and first request of each url.

Thus if during any given day a particular url is requested more than once then the time only of the last request is available in the global history record. In practice this is not a problem since the data analysis, which is described below, is based on *sessions* of Web information seeking activity which comprise each day's activity. There is no investigation of the duration of each session or, for example, the time intervals between elements of Web seeking activity.

As seen in Chapter two, discovering individual session boundaries is not generally possible and unobtrusively obtaining session durations is intractable; there can be an indicator for the start of a session when but without an explicit *logoff* it is not possible to say when a user concludes Web information seeking. Even in systems which do support explicit logoff the user may have concluded information seeking activity any time prior to logging off. Therefore in this investigation there is no attempt to temporally locate any information seeking more precisely than occurring within a daily session.[5]

In practice inspection of the data (together with some personal observation of student-users) suggests that repeat 'sub-sessions' during a single day are very unusual. This anecdotal conclusion is supported by the preponderance of small session click rates (see page 64) and a preliminary analysis of the Web request timestamps which showed that clicking in high click rate sessions seemed to be continuous or without an inter 'sub-session' break. Hence the daily definition of session is operationally equivalent to defining session using a more obtrusive basis.

The primary data collection survey procedure thus comprises two parts. Firstly there is a nightly global history file archiving component and then secondly there is the filtering and decumulation component which together produces the *Web log* file. Each component is now described under the respective headings, of i.) managing the data archive, and ii.) building the Web log.

i.) managing the data archive

Within each student-user's allocation of central storage space the global history file has the filename `netscape.hst`. Each night the automated copying procedure copies each `netscape.hst` file to a target file `yy_ddddddd.ext` where `yy_ddddddd` is an encryption of the student-user's student-id (which is also of the form `yy_ddddddd`). The student-user's year of registration prefix (or cohort) such as 98 is retained (in clear) and, most importantly, the encryption is consistent and unique so that each student-user is anonymously identified throughout the entire survey period. The

[5] The calendar mechanism was tested at the outset of the study to ensure that the internal history file clock was set to GMT/BST, rather than, say, PDT.

target file extension is of the form ddd which indicates the day of the survey. This starts at 000 on 10th October 1998 and increments by one each day. Hence all the target files copied from student-users in respect of a particular survey day all have the same file extension. Not all registered students use the Web so the number of global history files copied each night is less than the size of the sample-frame. (A file selection procedure to copy only those files which had been changed since the previous night failed part way through the survey after which every global history file was collected.[6])

Therefore each night there is a collection of typically many hundred individual student-user global history files each identified by the cohort, anonymous encrypted student-id and survey day. All of these files are compressed into a single global history archive (*ghar*) file with the name `yymmdd.zip` where this corresponds to the calendar day of the survey. Periodically these ghar files are transferred from the institution to the researcher's workstation to be processed in order to build the Web log.

ii.) building the Web log

Most generally a transaction log is a (computerised) list which records information relating to a transaction or some interaction between a user and a (computer) system. Peters, Kurth, Flaherty, Sandore & Kaske (1993a) describe the characteristics of a transaction log in the context of an OPAC which should include;

1. the characters input by the user,
2. a terminal identifier,
3. selected aspects of the system response, and
4. a *timestamp* of the transaction.

Here the Web log, which is constructed from the individual global history files contained in the ghar files, is for each student-user a complete list of every Web request (that is, visit to a Website) which he (or she) made during the two study-years. As well as the student-users and the Websites visited, the Web log also contains a record of session day and the number of visits to each Website. Each Web log record is,

<student-user identifier><session code><number of visits to Website><Website visited><?>.

[6] The selection procedure relied on interrogating the file date which it is believed was being corrupted to be the current date as a result of an unconnected system change.

The construction procedure includes both decumulating each student-user's global history files and filtering unwanted survey records. Each nightly ghar file when decompressed is a collection of (anonymous) individually named and date coded global history files. Each global history file is a BerkeleyDB (version 1.85) (Olson, Bostic & Seltzer, 1999) proprietary database in the form of an associative array (hash) or list of unique keys and their associated key-values (Bowers & Taylor, 2000; Christiansen & Torkington, 1998). The hash can be read using the Perl[7] DB_File package as in the Perl code fragment[8] which is illustrated in Figure 3.1.

```
use DB_File;
.
.

tie %hash, "DB_File", "global_history_file";
.
.

foreach $key ( keys %hash ) {
    $value = $hash{ $key };

    # do something with $key, $value record pairs
}
```

Figure 3.1: Perl code fragment to read a BerkeleyDB hash

Each unique hash key is a *url-string* that is the key is a text string which represents the url of the Internet resource which has been accessed. The hash value pointed at by the key is a sixteen byte encoded global history value which may have a text string of arbitrary length appended to it. This text string is the *html* title if it is present in the Internet resource. (In practice processing the global history files was an exercise in reverse engineering since no specification of their construction is available. It was found for example that the deletion of key-value record pairs from the hash was sometimes only partial and in consequence the standard Perl module (Bowers & Taylor, 2000) to read the history files was not stable.[9] I therefore wrote a more robust Perl program which also included audit procedures to verify that all records had been accounted for.)

The sixteen byte component of the global history record comprises four fields of four bytes each where each field is a binary packed long integer. The four integer fields

[7] Practical Extraction and Report Language, or "Pathologically Eclectic Rubbish Lister", (Wall, Christiansen & Schwartz, 1996).

[8] Difficulty was experienced following an operating system distribution upgrade because the current version of BerkeleyDB in the distribution, and hence DB_File is not compatible with version 1.85. This was overcome by reverting to the previously distributed version of Perl.

[9] This is believed to be related to the MS-Windows environment in which the history files were created; history files generated by Netscape/BerkeleyDB in a Unix like environment appear to be well behaved (Cockburn & McKenzie, 2001).

of the global history record are;

1. *last-time* or time of the last visit to the Website represented by the url-string the time in *epoch* format,
2. *first-time* or time of the first visit to the Website represented by the url-string in epoch format,
3. *visit count* of the cumulative number of visits to the Website starting from the time of the first visit, and
4. *expire flag* an internal software flag which distinguishes the records of files which are expressly requested (expire = 1) by the user from those of files such as embedded image files which are requested by the browser only because they are referenced in an html file. Image files which are expressly requested have expire = 1.

The epoch time format is the number of seconds which have elapsed since midnight 31 December 1969, hence for example, the global history record key value pair,

```
http://www.bris.ac.uk/ => 958826700 952262100 20 1
```

means that the url[10] <url:http://www.bris.ac.uk/> has been expressly requested by the user twenty times between epoch 952262100 and epoch 958826700, that is between 14:15pm on the 5th March 2000 and 13:45pm on the 20th May 2000.

As each student-user's global history file is read, the day of the last visit for each record (20th May 2000 in the above example) is compared with the encoded date of the survey day given by the file extension (which is also validated by the ghar file calendar name). If the record is *current*, that is the Website has been last visited during the survey day, and the expire flag indicates that the file has been expressly requested, then the global history record (that is key-value pair) is copied to a cumulative transaction log file which is maintained for each student-user.

Hence, after all the ghar files have been processed there is for each student-user a transaction log file which contains successive accumulated visit counts in respect of a url-string for every url expressly requested by that student-user at any time throughout the two study-year survey period. These can be processed to decumulate the visit counts, for example the successive cumulative records,

[10] The textual rendering of uniform resource locators which is used here is as recommended in RFC 1738 (Berners-Lee, Masinter & McCahill, 1994).

`http://www.bris.ac.uk/ => 958826700 952262100 20 1`, and

`http://www.bris.ac.uk/ => 958923900 952262100 25 1`

which refer to the 20th and 21st May 2000 (but successive records are not necessarily successive days) imply that the url `<url:http://www.bris.ac.uk/>` was requested (expressly) five times during the 21st May 2000 daily survey.

Table 3.1 illustrates an example extract from the Web log which is produced by processing, including decumulating the visit counts, student-users' individual cumulative transaction log files. Each line in the extract is a Web log record which as mentioned above has four components,

1. *user-id* which is the ten character anonymizing encrypted student-id (in the example the student-user is from the 1997 cohort and the three trailing digits are intentionally hidden),

2. *session* which is the concatenated year and day number of the day of the survey (24 May 1999 in the example),

3. *click frequency* the decumulated visit count (during the session) which corresponds to the *conditioned url-string*, and

4. *conditioned url-string* derived from the url-string recorded by the global history survey of Web information seeking activity. The derivation of the conditioned url-string is given below.

```
97x0000*** 1999144 1 www.yahoo.akadns.net/
97x0000*** 1999144 1 search.snv.yahoo.com/bin/search?
97x0000*** 1999144 1 search.snv.yahoo.com/search?
97x0000*** 1999144 1 dir.yahoo.akadns.net/Arts/Humanities/Literature/Genres/
97x0000*** 1999144 1 search.snv.yahoo.com/bin/search?
97x0000*** 1999144 1 dir.yahoo.akadns.net/Entertainment/Music/Artists/By_Genre/Industrial/Industrial_Dance/Sheep_on_Drugs/
97x0000*** 1999144 1 dir.yahoo.akadns.net/Entertainment/
97x0000*** 1999144 1 dir.yahoo.akadns.net/Entertainment/Music/Artists/By_Genre/
97x0000*** 1999144 1 dir.yahoo.akadns.net/Entertainment/Music/Artists/By_Genre/Rock_and_Pop/Surf_Rock/
97x0000*** 1999144 1 angelfire.com/biz/angle3/
97x0000*** 1999144 1 angelfire.com/biz/angle3/confront.html
97x0000*** 1999144 1 www.jukebox.demon.co.uk/
97x0000*** 1999144 2 dir.yahoo.akadns.net/Arts/Humanities/Literature/
97x0000*** 1999144 2 dir.yahoo.akadns.net/Arts/Humanities/Literature/Genres/Science_Fiction_and_Fantasy/Authors/Huxley__Aldous/
97x0000*** 1999144 1 www.primenet.com/ matthew/huxley/
97x0000*** 1999144 1 www.primenet.com/ matthew/huxley/ThemesInHuxley.html
97x0000*** 1999144 1 www.cyber-nation.com/victory/quotations/authors/quotes_huxley_aldous.html
97x0000*** 1999144 1 home.cp-tel.net/miller/BilLee/quotes/Huxley.html
```

Table 3.1: Extract from the Web log

The url-strings in the global history survey, each represent a unique[11] Internet resource but are not of themselves unique. That is, each url-string can have many equivalent but different string representations. Since the essence of this investigation is to consider student-users' Web information seeking activity on the basis of the Websites which they visit, then a reliable procedure is required to identify Websites. Just comparing url-strings is not sufficient since, to begin with;

`<http://foo.com>` and `<http://foo.com/>` are the same,

`<http://FOO.COM>` and `<http://foo.com>` are the same,

`<http://foo.com/index.html>` is the same as `<http://foo.com/>`, and

`<http://foo.com/>` is probably the same as `<http://foo.com/index.html>`,

`<http://foo.com/>` is also probably the same as `<http://www.foo.com/>`,

and may be the same as an alias `<http://aka-foo.com/>`, and

`<http://foo.com/>` will be the same as one or more dotted decimal *Internet Protocol* (IP) address[12] `<http://nnn.nnn.nnn.nnn/>`.

A *conditioning* procedure has therefore been devised[13] which converts each url-string into its conditioned url-string which is then a more reliable representation for each Website. This is a multistage procedure involving four steps.

Conditioning – step one: Canonically each url is of the form (Berners-Lee *et al.*, 1994),

<scheme>://<user-name>@<host-name>:<port-number>/<path>#<internal-anchor>?<search-part>

which, in the absence of a <user-name> or <port-number>component, is equivalent to the form <scheme>://<full-path>#<internal-anchor>?<search-part>.

Hence any canonically invalid url-strings, say `<htp://www_foo..comm/bar/in valid>`[14] can be identified and are immediately discarded.

[11] In the sense of being a single particular file or assemblage of files. Identifying *mirrored* Internet resources or copies of the file elsewhere in the Internet is discussed by Bharat, Broder, Dean & Henzinger (2000) and Cothey (2001).

[12] The machine readable IP address is a sequence of four binary numbers each in the (decimal) range 0 to 255. When these are expressed as, say, 137.222.10.46 for human consumption the IP address is said to be in dotted decimal form.

[13] This can be thought of as analogous to the stemming procedure used in IR (Porter, 1980).

[14] The scheme, host-name format, top level domain and character set, and path character set all invalid.

The *host-name* must comply with an overall structure but within that structure considerable variation is possible. By default host-names are in dotted decimal IP form. However host-names conforming to a naming convention and format can also be constructed from a (restricted) set of alphanumeric characters. When an alphanumeric host-name is used there must be an associated DNS registration which facilitates the conversion to IP address. The DNS allows host-name aliases so that both <foo.com> and <www.foo.com> may be registered to the same IP address(es) and thus access the same Internet resource. In order to maintain control, each different Internet host (server computer) in the DNS has a unique *official name* to which all the DNS aliases relating to that name refer. The DNS can be interrogated to return the IP addresses and official name associated with a valid registered by host-name. Host-names which are invalid or not registered cause a null response. The DNS does not distinguish between upper and lower case alphabets, so that <FOO.com> is regarded as being same as <foo.com>. If a host-name in dotted decimal form also has an alphanumeric host-name then the official name for this can also be found from the DNS.

The first step to condition the url-string verifies that each url-string is canonically valid and to discards any user-name or port-number (and associated @ and : delimiters). The DNS is next interrogated with each canonically valid host-name and if an official name is available then the host-name in the conditioned url-string is replaced by the official name. The residual host-name cases are those host-names which are either in dotted decimal IP form or in a valid alphanumeric host-name form but do not have an associated DNS entry. For valid alphanumeric host-names this usually indicates a DNS registration change between when the Web request survey was carried out and when the DNS is being interrogated. The host-name in each conditioned url-string is then munged by being replaced by either its official name in lower case or its residual alphanumeric name in lower case but where any leading <www.> is removed.

Conditioning - step two: After step one each partially conditioned url-string is canonically of the form, <scheme>://<full-path>#<internal-anchor>?<search-part>.

Unwanted url-strings which represent schemes (such as <file> or <telnet>) where the associated activity falls outside the scope of Web information seeking are ignored. Two other categories of url-string are also filtered out. These are;

1. Url-strings in respect of the institution's own Web resources, intranet and library catalogue. This is because some of the institution's own teaching and learning material is delivered electronically. It was considered that if this material were included then the survey would be distorted by, for example, the

anticipated expansion of such material between the two study-years and any changes in the proportions of student-users undertaking the electronically supported courses.

2. Email and Chat-rooms. During the survey the method of access to the student email service changed to be by way of a Web based interface. This together with the development of Web based email services generally potentially distorts the survey and therefore whenever activity which is just gaining access to an email account is detected then this is filtered out. Chat-room access is treated similarly. The overarching rationale is that neither activity is regarded as part of how a user locates Web information.

The scheme, scheme delimiter, search-part and internal-anchor are removed. The internal-anchor delimiter, #, is removed later but the search-part delimiter, ?, is retained.

Conditioning - step three: Thus after step two each partially conditioned url-string should be canonically equivalent to, <full-path> together with a possible ? termination. Therefore any bare host-names such as `<foo.com>` are replaced by the properly terminated full-path `<foo.com/>` and the default files in paths, such as `<index.html>` are removed.

Conditioning - step four: The final step of conditioning the url-string in order to provide a reliable indicator of Website is to adjust the click frequency within each session to aggregate clicks to the same Website. For example, suppose that during a particular session a student-user's embryonic Web log contained before conditioning;

```
2 foo.com/

1 foo.com/index.html
```

where the numbers are the decumulated click frequencies, then after conditioning this would be;

```
2 foo.com/

1 foo.com/
```

so that the Web log entry is aggregated to;

```
3 foo.com/
```

Greater complexity arises in respect of conditioned url-strings which, at this step, contain either (or both) of # and ? showing that the url-string represented a url containing an internal-anchor or search-part respectively.

Usually the search-part is unique therefore the click frequency of, say **<foo.com/search?>** would be one but within any session there may be several such instances. These are not aggregated (see example illustrated in Figure 3.1). However if the search-parts are identical (within a student-user's session) then the click frequencies are adjusted as described above.

The phenomenon of internal-anchors is not generally recognised and is referred to only by Tauscher (1996, p. 52)[15] Link-clicking on internal-anchors provides just a mechanism to scroll through a Website so there are two issues. Firstly two users may be distinguished only by whether or not they link-clicked an internal-anchor and secondly, a Website may change by only the provision of an internal-anchor. When distinguishing two users one of whom visits a Website twice the other only once, the internal-anchor activity is an artifact. One user only *appears* to have visited more than the other although the Web information located is identical (because a Web page is taken to be of arbitrary length). This applies also in the case of a user revisiting a Website after an internal-anchor has been inserted (or removed). Hence, in general, click frequency is not incremented by the presence of internal-anchors. There are two cases. Suppose;

```
2 foo.com/bar.html

1 foo.com/bar.html#

3 foo.com/bar.html#
```

then the click frequency is taken to be 2 and the two **<foo.com/bar.html#>** url-strings each representing a link-click to a different internal-anchor are ignored, and secondly,

```
1 foo.com/bar.html#

3 foo.com/bar.html#
```

[15] Tauscher truncates internal-anchors but her algorithm is much simpler than that used here because she used software instrumentation to monitor *every* url request.

when the click frequency is taken to be 1, that is the minimum click frequency of the two internal-anchors based Web requests.

Both # and ? can occur in the partially conditioned url-string, say `<foo.com/bar.html#search?>`. This is converted to `<foo.com/bar.html?>` and then treated as described above depending on the associated search-part.

Independently of the devising of this survey and data collection procedure McKenzie & Cockburn (2001) published a survey procedure which appears to be almost equivalent. They correct themselves in Cockburn & McKenzie (2001, p. 907) since they "incorrectly excluded [Websites] with a variety of suffixes, including `.jpg` and `.gif`." Their later procedure used the expire flag as used here. They also "normalised" urls with search-parts (by removing them) but make no mention of internal-anchors or Webhost aliases.

3.2.3 Data analysis

The principal primary data of the investigation are the conditioned observational records of the Web log each of which comprises, user-id, session, click frequency, and conditioned url-string. An extract from the Web log is illustrated in Table 3.1 on page 57.

The data analysis described here is the extrinsic transaction log analytic and the meta-analytic features of the investigation. Intrinsic analysis of the Web log which is based on an interpretation and analysis of the Website content *within* each Web log record is discussed below in Section 3.3.

TLA consists of identifying features of interest in the transaction log, counting their frequency of occurrence and possibly relating the incidence of one feature with another. Transaction logs offer a much richer variety of features than are relevant to the research questions in hand. Hence particular features must be selected. These must be assessed in respect of both reliability and validity, (for example the reliability of the character string representation of the Website has been discussed above). The misinterpretation in the hypothetical OPAC TLA investigation discussed in Chapter one results from a lack of validity.

A transaction log is usually a snapshot survey of activity during a short period. The unit of analysis is often a *session* of one or more transaction records from a single user but individual users are not usually identified from session to session. Analyses *by-session* are relatively straightforward to undertake but analyses *by-user* are not generally possible. Both for this reason and because of the short duration of snapshot, how individual users change over time cannot be discovered.

The transaction log analysed here consists of two extended surveys (each study-year) within which each user is individually (anonymously) identified. Each extended survey is a longitudinal study in its own right one of which has been reported (Cothey, 2002). Since users are identified and the identification is consistent then by-user analysis can be carried out and how Web information is located by an individual user can be compared between study-years.

The by-user data analysis focuses on each individual user in the transaction log and constructs as secondary data a characterization of each user based on features in the transaction log. For example a simple Perl script composed by the investigator scans the Web log counting users and sessions and reports for each user the number of sessions undertaken during each of the study-years or the user's *session rate.* Perl is designed around processing strings of characters and is especially suited to the computationally intensive demands of TLA. The Web log consists of 1,050 student-users who collectively undertake 46,558 sessions comprising 1,990,488 clicks or requests for Web information. An eligibility criterion of two sessions or more of Web information seeking during each study-year is set in order that a *trajectory analysis* (discussed below) can be carried out.

The data analysis is influenced by the heavy tail phenomenon mentioned in Chapter two. Huberman *et al.* comment is respect of their data, that:

> This distribution has two characteristics worth stressing in the context of user surfing patterns. First, it has a very long tail, which extends much farther than that of a normal distribution with comparable mean and variance. This implies a finite probability for events that would be unlikely if described by a normal distribution. Consequently, large deviations from the average number of user clicks computed at a site will be observed. Second, because of the asymmetry of the distribution function, the typical behavior of users will not be the same as their average behavior. Thus, because the mode [sic] is lower than the mean, care must be exercised with available data on the average number of clicks, as this average overestimates the typical depth being surfed.
>
> (Huberman *et al.*, 1998, pp. 95)

The data of this investigates exhibits the same characteristics. Thus Figures 3.2 and 3.3 shown below illustrate a more power law than Gaussian appearance to the frequency distribution of both the by-user session rate and by-session session click rate for student-users where session click rate is the click frequency during a session. The secondary data is dominated by the constraints imposed by the session click rate distribution since a large proportion of daily Web information seeking sessions

involve only a few clicks. In consequence of this distortion the inference procedures discussed below make use of nonparametric statistical techniques (Siegel & Castellan, 1988).

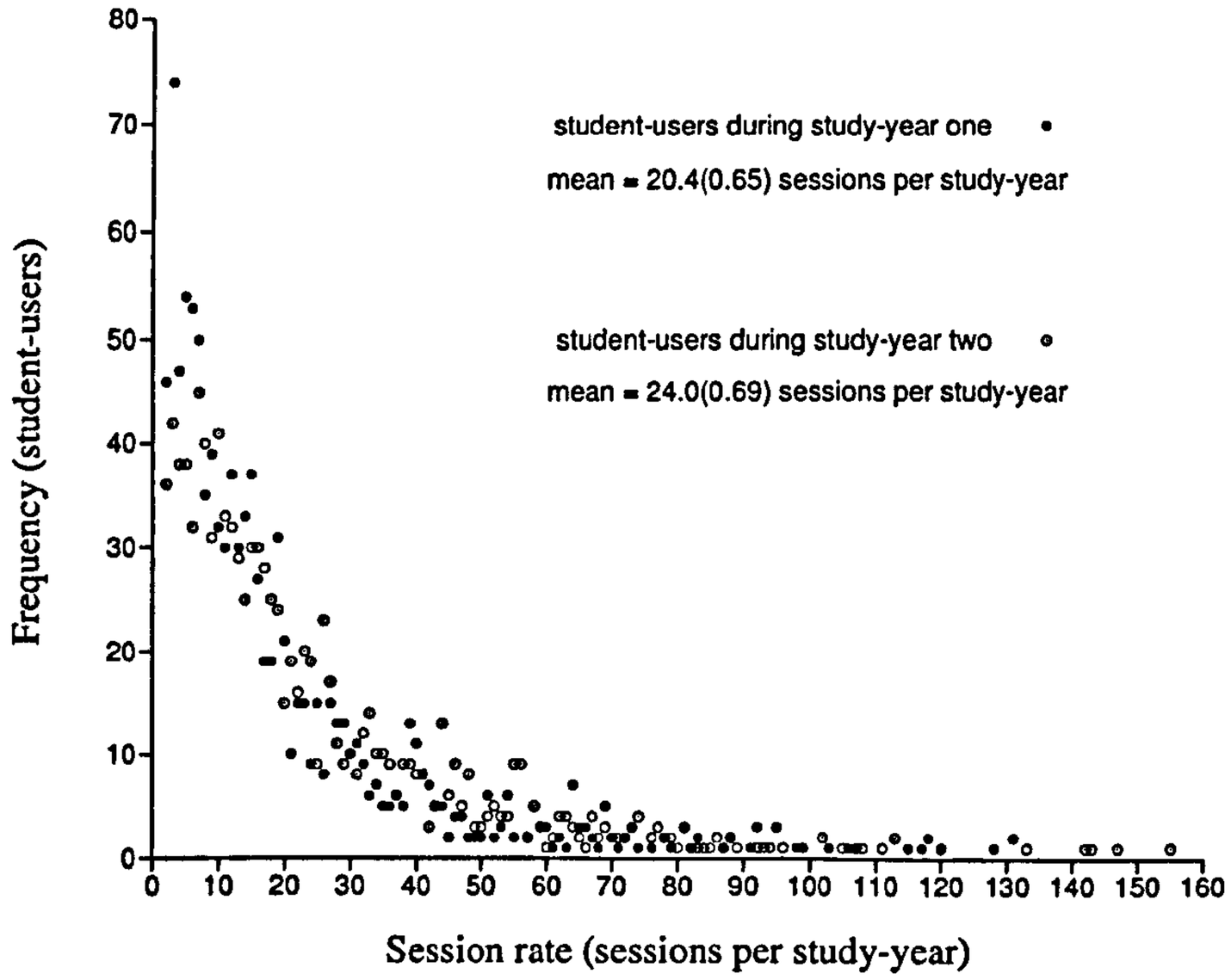

Figure 3.2: Frequency distributions of student-user's session rate

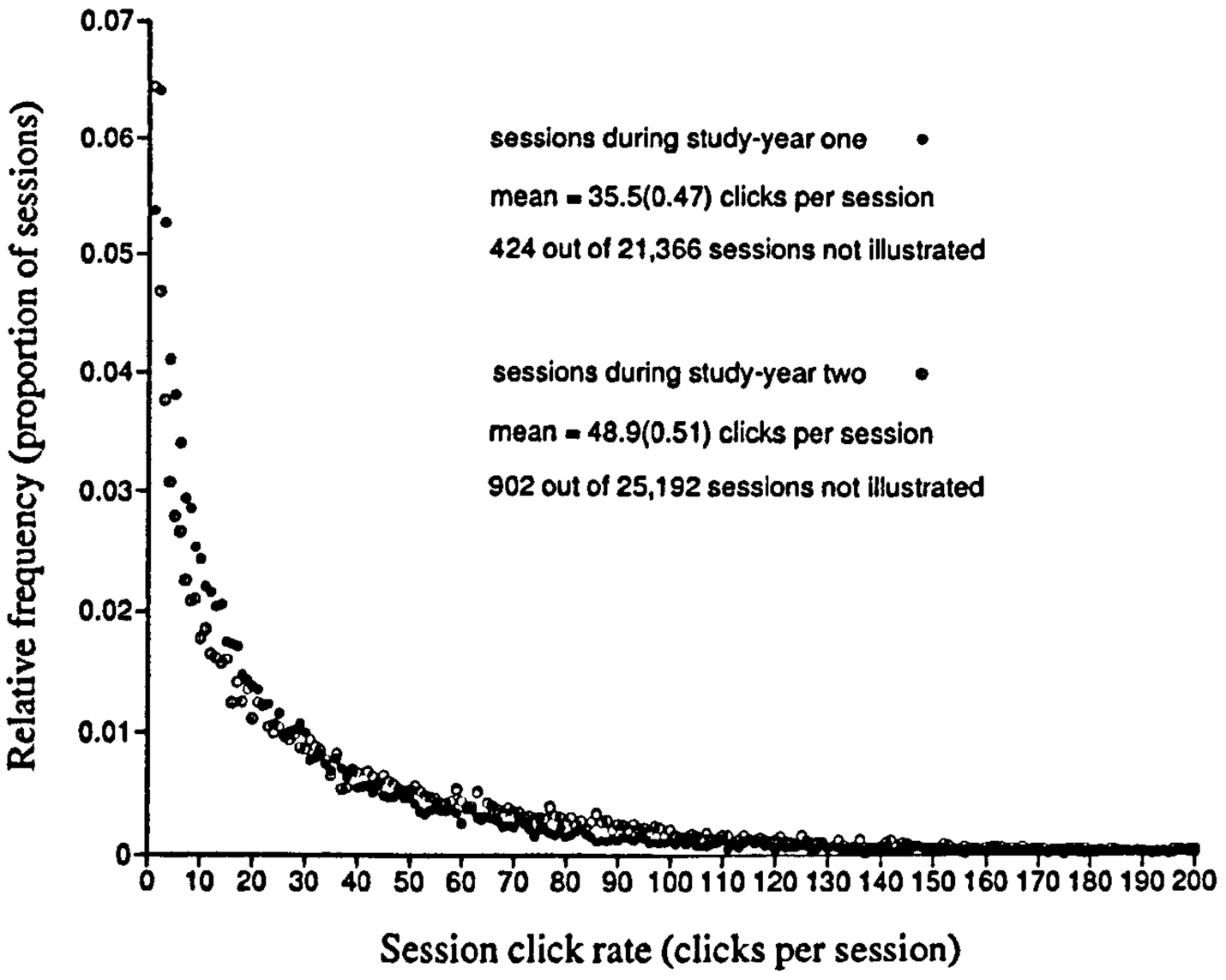

Figure 3.3: Relative frequency distributions of session click rate (range illustrated up to 200 clicks per session)

A simple qualitization (Tashakkori & Teddlie, 1998) procedure characterises student-users as having either a *small* or *large* session rate depending on whether or not their session rate during the study-year is less than or greater than the mean session rate. This qualitized data is then used meta-analytically to interpret the derived secondary data. The secondary data is distorted by most student-users having small session rates and most sessions containing few (compared to the mean) clicks.

The median session rates (by-user) for study-years one and two are about 13 and 16 sessions per study-year respectively. An alternative qualitization basis would be to use these median values in place of the mean. Because of the shape and heavy-tail nature of the distribution (that is, the preponderance of small session rates combined with a few very large session rates) the investigation's findings are not affected by which qualitization criterion is used. However since a qualitization based on the mean was simpler to implement as regards the variety of analyses undertaken then this is preferred over a median based qualitization.

The meta-analysis is discussed more fully in Section 3.3 where the metrics used to describe and characterize each how each student-user locates Web information are described. An example of meta-analysis is the analysis of session click rate. This shows that in respect of the earlier extended survey (study-year one) the mean session click rate is 35.5(0.47) clicks per session (the mean is 35.5 and has a standard error of 0.47) while during the later survey (study-year two) the mean session click rate 48.9(0.51) clicks per session.

The z test of difference which is discussed in the next Section shows that this increase is more than might be explained by random variation and hence one might infer that student-users' Web information seeking involves more clicks during study-year two than during study-year one. There could be several explanations for this (which are not exclusive). For example the duration of student-users sessions during study-year two may be longer, there may be structural changes in the Web which make information more *diffuse* so that two clicks are required where previously one sufficed, or student-users may really be more energetic (say in clicks per minute). The inference is therefore ambiguous; disambiguation by comparing the two study-years is discussed in the next Section.

Longitudinal investigations study phenomena over time and a wide range of phenomena including the Web are studied in this way (for example, Cooper, 2001; Koehler, 2002). However longitudinal studies of information behaviour are uncommon (Yuan, 1997).

When the longitudinal method involves repeated measures over time on the *same* individual the design is called *longitudinal-developmental* (Nesselroade & Baltes, 1979)

in order to focus attention onto the by-individual nature of the study. Longitudinal-developmental analysis and in particular conditional analysis which investigates the existence of change is discussed in Section 3.6.

3.2.4 Inference procedure

The inference procedure is based mainly on statistical hypothesis testing the results of which are then collated in a more qualitative fashion to construct descriptive narrative profiles. Most of the secondary data are in the form of an SPSS (SPSS, 1999) dataset which facilitates generating descriptive statistics. This dataset is described in Section 3.3.

Both parametric and nonparametric testing procedures are used. The rationale and application of the tests used are illustrated below. The gender of each student-user is determined from the linked demographic data.

z test of difference (Kanji (1999) lists this as Test #3.) For example in respect of the session rate data shown in Figure 3.2 the rationale is that the means of $\bar{x} = 20.4(0.65)$ and $\bar{y} = 24.0(0.69)$ sessions per study-year for study-years one and two respectively are each distributed normally (central limit theorem) even though the underlying distribution may not be normal. Therefore, $\bar{x} - \bar{y}$ is normally distributed with mean $\mu_x - \mu_y$ and variance $\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$ where μ and σ^2 are the population mean and variance, and n is the sample size (n is large and is mostly 1,050).

Hence if it is assumed as a null hypothesis that $\mu_x - \mu_y = 0$ then the test-statistic

$$z = \frac{20.4 - 24.0}{\sqrt{0.65^2 + 0.69^2}} = -3.8$$

is, by hypothesis, drawn from the standard normal distribution.

The region of rejection may by one or two sided depending on how the alternative hypothesis is formulated and the critical value depends on the value of α (or level of significance) which is chosen. The conjecture that the session rate during study-year two is greater than during study-year one is tested at $\alpha = 0.01$ by reference to the critical value $z_{0.01} = 2.33$. Since $z = 3.8 > z_{0.01}$ the null hypothesis is rejected and the alternative is accepted. This is reported as, $24.0(0.69) > 20.4(0.65)$ ($p < .01$, $z = 3.8$).

χ^2 tests (Kanji (1999) lists these as Tests #37, #40 and #44.) Three similar χ^2 nonparametric tests are used, goodness of fit, consistency and independence.

To test whether or not the gender mix of the student-users conforms to the mix in the sample-frame then a goodness of fit test is used. This is illustrated in respect of the 1997 cohort by Table 3.2 which shows an observed frequency of 203 men and 219 women student-users compared to a theoretical expected frequency of 181.4 men and 240.6 women based on the gender mix of the sample-frame.

Distribution	**Student gender**		**Both**
	men	women	
Sample (student-users) observed frequency	203 students	219 students	422 students
Theoretical expected frequency	181.4 students	240.6 students	422 students
sample-frame	904 students	1,199 students	2,103 students

Table 3.2: Gender 'goodness of fit' student-user frequencies for the 1997 cohort

The test-statistic is

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected frequencies respectively, which is distributed as χ^2 with one degree of freedom. Thus

$$\chi^2 = \frac{21.6^2}{181.4} + \frac{21.6^2}{240.6} = 4.51$$

and exceeds the one-sided critical value for χ^2 with one degree of freedom which for $\alpha = 0.05$ is 3.84. Hence the 1997 cohort sample gender mix is not a good fit to the sample-frame. The goodness of fit test is reported as $\chi^2 = 4.51 > \chi^2_{1;0.05}$.

For the 1998 cohort, $\chi^2 = 6.12$ so that the 1998 cohort sample gender mix is also not a good fit to the sample-frame. In each case, inspection of the observed and expected frequencies shows a male bias.

The χ^2 test of consistency investigates whether or not two samples are drawn from the same population. Using the data in Table 3.3 the test-statistic is

$$\chi^2 = \frac{1,049 \times (203 \times 291 - 337 \times 219)^2}{422 \times 540 \times 510 \times 628} = 3.12$$

Student-user cohort	Student-user gender		Both
	men	women	
1997 cohort	203 students	219 students	422 students
1998 cohort	337 students	291 students	628 students
Overall	540 students	510 students	1,050 students

Table 3.3: Gender 'consistency' student-user frequencies for the 1997 and 1998 cohorts

The test-statistic $\chi^2 = 3.12$ is compared with the critical value of χ^2 with one degree of freedom. At $\alpha = 0.05$, $\chi^2 = 3.12 < \chi^2_{1;0.05} = 3.84$ and the null hypothesis is not rejected. It is accepted that the two samples are consistent in that they can be drawn from a single underlying population (which therefore implies that the male bias just demonstrated is *consistent* or occurs during both study-years).

The third χ^2 test which is used is the χ^2 test of independence. This is illustrated by the cross-tabulation of sessions shown in Table 3.4. The frequencies are of sessions during study-year two each classified by two attributes, the cohort of the originating student-user and a *smaller/larger* qualitization of the session click rate when compared to the mean session click rate, see Figure 3.3.

Student-user cohort	Study-year two click rate		Both
	smaller	larger	
1997 cohort	6,521 sessions	3,314 sessions	9,835 sessions
1998 cohort	10,823 sessions	4,534 sessions	15,357 sessions
Study-year two	17,344 sessions	7,848 sessions	25,192 sessions

Table 3.4: Cohort and click rate 'independence' during study-year two

Based on the marginal frequencies, the expected frequencies are,

$$\frac{9,835 \times 17,344}{25,192} = 6,771$$

$$\frac{9,835 \times 7,848}{25,192} = 3,064$$

$$\frac{15,357 \times 17,344}{25,192} = 10,573$$

$$\frac{15,357 \times 7,848}{25,192} = 4,784.$$

The test-statistic is

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected frequencies respectively, which is distributed as χ^2 with one degree of freedom. Thus

$$\chi^2 = \frac{250^2}{6,771} + \frac{250^2}{3,064} + \frac{250^2}{10,573} + \frac{250^2}{4,784} = 48.6$$

which exceeds the critical value (at $\alpha = 0.005$) since $\chi^2_{1,0.005} = 6.63$ and hence the null hypothesis that the two attributes are independent of each other is rejected. This is reported as the cohort and click rate attributes are not independent ($p < .005,\ \chi^2 = 48.6 > \chi^2_{0.005} = 6.63$).

The inference procedure based on comparing the two study-years is illustrated by considering the relationship between the cohort and click rate attributes which has just been shown. During study-year one[16] the cohort and click rate attributes are also not independent ($p < .005, \chi^2 = 225 > \chi^2_{0.005} = 6.63$) and by inspection, during each study-year, sessions from the 1998 cohort of student-users are biased towards being *smaller* (have fewer clicks than average). Thus the conclusion that the sessions of student-users from the 1998 cohort are more likely to be *smaller* than are sessions of student-users from the 1997 cohort is reinforced because the conclusion is consistent for each study-year. If this were not the case then such a claim would invite further examination.

A similar argument helps to disambiguate the effects of structural change in the Web and change in how users locate Web information. In this case the relationship examined is across time and the consistency sought is between groups of users. Since it is considered that structural change should affect all users (although not necessarily uniformly) then a change over time which affects some users but not others (and in

[16] The session frequency cross-tabulation in respect of study-year one is given in Table B.1 in Appendix B.

the extreme, is reversed) is less likely to have a structural origin than a change over time which affects all users.

The discussion above also illustrates the use of a narrative profile to compare an aspect of the Web information seeking activity of student-users from the 1997 and 1998 cohorts. The investigation develops such profiles by conjecture and refutation. The later Sections of this Chapter describe how user characterization metrics are derived from the Web transaction log. These characterization metrics discriminate between user's Web information seeking activity and thereby refine the profiles.

Change over time is also discussed in Section 3.6.

3.2.5 Ethical issues

The research design follows the dominant ethical practice of large scale Web information behaviour research which is to adopt the permission model and to take positive steps to ensure the privacy of individual users, see Chapter two. Additional ethical guidance is also available from the principles which relate to data protection in the UK.

The discussion here therefore sets out the justification for believing that permission is authorised, the conditions of the authorisation and the additional research practice in excess of these conditions in order to ensure individual privacy.

The focus of attention is the global history file in the context of a student-id. The global history file is *owned* by the institution and if it contains *personal data* then it is subject to the Data Protection Acts (1984 subsequently 1998). In theory there should be no personal data present however it was deemed to be personal data and the advice of the Data Registrar was sought. The Acts provide that, for the purposes of research, a data-owner can release data to a researcher provided that individual confidentiality is maintained, no decisions are arrived at in respect of any individual to whom the data refers and that substantial damage or distress is not caused. The responsibility for interpreting the Act rests with the Court so that the Data Registrar cannot guarantee his (or her) advice, however the Court has already ruled favourably in respect of data released for the purposes of research. Hence the Data Registrar considered the advice to be well founded and the institution is reliably authorised to release the global history file data (and the demographic data) even though there is no prior data registration that the data may be used for research.

Permission was granted by the official within the institution who was responsible for ensuring the institution's compliance with the Data Protection Act. Hence the belief that there is a grant of authorised permission is justified.

Since the global history files are linked to the student-id which is itself linked to the student's name by way of a publicly available index, then additional research practice is followed to ensure the confidentiality of the data. This consists principally of an encrypted recoding of the student-id (Penniman & Dominick, 1980). The encryption is bijective and the decryption is available. (The decryption facility was introduced during the planning of the research in case of a need to interview student-users about their Web information seeking activity.) All the codes which can identify individuals in the research data are encrypted so that individual confidentiality is maintained and no individual's data revealed by the research would be passed back to the institution. It was also agreed that the research data would be published only in the form of grouped data and not as individual data.

As a final protection the research practice is to remove any identification which might inadvertently be present in the data. This can occur when names are entered as part of a url-string either as user-names or more usually incorrectly.[17]

The longer term custody of the research data archive is the responsibility of the investigator although it is understood that the data remains owned (in a data protection sense) by the originating institution.

3.3 Web log analysis

This Section describes the intrinsic Web specific analysis of the Web log which examines the conditioned url-strings in order to construct the secondary dataset which forms the principal basis of the investigation. The analysis of search-parts which is also carried out is described in Chapter five.

The Web log records for each student-user for each daily session of Web information seeking, the collection of conditioned url-strings and their associated click frequencies which represents how the student-user locates Web information. Each conditioned url-string is of the form, <Website><?> which is equivalent to, <Webhost>/<path><?> but where the terminating <?> is only sometimes present.

As described previously the Webhost string representation is conditioned or munged so that the Website reliably distinguishes requests by the student-user for different Internet resources. Different Websites are different in the lexical sense that they do not have the same conditioned url-string. The *substance* of the Internet resource addressed by the Website (were it to be reconstituted as a valid url) is not considered, thus it is possible that two different Websites are semantically identical. In such

[17] An example is accidently using the url field of the graphical browser to compose an email.

circumstances the resources are *mirrored.* The Websites correspond to the student-user's *requests* for Web information since *embedded* resources, such is image files, are excluded. Images per se are not excluded so that Websites representing expressly requested image files can occur in the Web transaction log.

The Website *vocabulary* is defined to be the set of different Websites, and similarly the Webhost vocabulary is the set of different Webhosts. Vocabulary sets will correspond to different time periods so that there is a *session vocabulary* and a *study-year vocabulary.* The cardinality of the set is referred to here as the *repertoire.*[18] Hence;

$$\sum \text{clicks} \geqslant \text{Website repertoire}$$

$$\text{Website repertoire} \geqslant \text{Webhost repertoire}$$

and;

$$\sum \text{session repertoire} \geqslant \text{study-year repertoire.}$$

The terminating query, <?>, indicates that a search-part has been submitted to the Website or more precisely, to the resource which is addressed by the valid url reconstituted from the Website character string. From now, for brevity, this qualification is understood. Hence the Web log distinguishes two forms of information seeking activity, link-clicking and query-clicking. Query-clicking indicates a more active form of information seeking compared to link-clicking since particular user supplied information is provided. (It is accepted that urls which are manually keyed link-clicks require additional keystrokes from the user; they do not however require any user data.)

Analysis of the Web log is in two stages (each of which requiring the investigator to compose large Perl scripts). Firstly each student-user's sessions are analysed and summarised to obtain, for each session, data of the form,

<user-id><session code><session click rate><session query-click rate><session Website-repertoire>
<session Webhost-repertoire><cumulating study-year Website-repertoire><cumulating study-year Webhost-repertoire>

where,

<user-id> is the ten character anonymized student-id,

[18] Other authors use just "vocabulary" or "vocabulary size" (Greenberg, 1993; Thomas, 1998).

<session code> is the seven character concatenated year and day number,

<session click rate> is the total click frequency during the session which is the sum of the click frequencies in the Web log,

<session query-click rate> is the total query click frequency during the session which is the sum of the query click frequencies in the Web log,

<session Website-repertoire> is the number of different Websites visited during the session,

<session Webhost-repertoire> is the number of different Webhosts visited during the session,

<cumulating Website-repertoire> is the cumulative number of different Websites visited during the study-year, and

<cumulating Webhost-repertoire> is the cumulative number of different Webhosts visited during the study-year.

For example, using the extract from Table 3.1, the first stage of the secondary data is,

97x0000*** 1999144 20 3 17 9 <Website study-year repertoire><Webhost study-year repertoire>

which illustrates that two of the queried Websites are the same.

There are two logical by-session datasets, one for each of study-year one and study-year two, which contain 21,366 and 25,192 session records respectively.

The second stage is to summarise these datasets to obtain datasets of the form,

<user-id><session rate><∑session click rate><∑session query-click rate><∑session Website-repertoire><∑session Webhost-repertoire><study-year Website-repertoire><study-year Webhost-repertoire>

for each study-year.

The findings presented in Chapter four make use of seven user-characterization metrics which describe and differentiate how student-users locate Web information. Six of these metrics are intrinsic or based on interpreting the *content* of the Web log record. The rationale and definition of each is given in Chapter four in the context of developing the narrative profiles already mentioned. Four of the metrics, for

example the *average Webhost-persistence* which is derived as described below are obtained directly from the datasets described above. The example of this metric is chosen only to demonstrate the analytic method. The analyses required to obtain the three *trajectory* and *conformance* based metrics is described later in the next two Sections.

The average Webhost-persistence for each student-user during each study-year is;

$$\frac{\sum \text{session Website-repertoire}}{\sum \text{session Webhost-repertoire}}$$

This is the number of Websites visited at each Webhost and is measured in Websites per Webhost.

For each session, the ratio $\frac{\text{session Website-repertoire}}{\text{session Webhost-repertoire}}$ gives the average number of different Websites visited within each (different) Webhost. This is similar in principle to the *path length* used by Huberman *et al.* (1998). The metrics are not identical because Huberman *et al.*'s Law of Surfing analysis starts a new session each time a different Webhost is visited. Therefore the user's visit sequence;

```
<url:http://foo.com/>
<url:http://foo.com/one/>
<url:http://foo.com/two/>
<url:http://foo.com/one/>
<url:http://bar.com/one/>
<url:http://foo.com/>
```

which is here analyzed as session Webhost-persistence $= \frac{4}{2} = 2$, is three paths of length three, one and one and hence an average path length of 1.7 for the Law of Surfing. Although the Law of Surfing is computationally sophisticated, the path length metric is affected by local caching so that the example could just as well be two paths of length three and one (if the second request for `<url:http://foo.com/>` were satisfied by the user's local browser cache[19]). Huberman *et al.* do not report any conditioning of urls other than the exclusion of embedded files therefore the path length metric double-counts internal-anchors and thus overestimates path length compared to session Webhost-persistence. The aggregate data which they report is "3,247,054 page requests from 1,090,168 Web sites" (for page read Website and for Web site read Webhost) hence their overall Webhost-persistence is $\frac{3{,}247{,}054}{1{,}090{,}168} = 2.98$.

[19] The data were collected from a proxy-server log.

The average Webhost-persistence for each student-user which is used here is a weighted mean of the student-user's session Webhost-persistence which estimates the typical Webhost-persistence of the student-user. As the session Webhost-repertoire falls or as the interleaving of Webhosts in a session reduces so Webhost-persistence and path length will become equivalent (after adjusting for caching and internal-anchor effects).

3.4 Web vocabulary and trajectory analysis

The idea of a *Web vocabulary* has been used both explicitly and implicitly in previous research, see Chapter two. The Website-vocabulary, for example, is a set of different Websites. The Website-repertoire is the cardinality of the Website-vocabulary.

The explicit use of Web vocabulary and repertoire was pioneered by the OCEANS group who used a fractal random walk mathematical model of repertoire growth. This contrasts with the informal approach of the Greenberg school. Both groups focus on the slope of the repertoire growth curve or *trajectory* as a metric which characterizes Web information seeking although the detail differs. The OCEANS group transform their data logarithmically before calculating the slope but the Greenberg school do not. Since neither provide any evidence in support of their method then the alternatives are examined in order to decide which method to adopt.

Suppose the trajectory function, τ say, is,

$$\text{repertoire} = \tau(\text{clicks})$$

which shows how the vocabulary of a user develops as the number of clicks by the user increases. Then at the extremes, a user who never revisits any Website will have a Website-trajectory function $\tau(n) = n$ while a user who always visits the same Website will have a Website-trajectory function $\tau(n) = 1$. The random walk model implies that,

$$\tau(n) \propto n^{\theta} \text{ where } 0 \leqslant \theta \leqslant 1$$

which is as expected from Heaps' Law (Heaps, 1978).[20] The value of θ (or so called fractal dimension) characterizes that user's Web information seeking activity.

Since,

$$\text{Ln}(\tau(n)) = \theta \text{Ln}(n) + \text{constant}$$

[20] None of the literature from Computer Science makes any reference to Heaps' Law which on pragmatic grounds suggests that a power law should apply.

then θ is the slope of the (least squares best fit) straight line fitted to the logarithms of the observed click and repertoire data.

Using such a procedure as done by the OCEANS group tacitly assumes that actual repertoire growth conforms to the random walk (or Heaps' Law) better than to an alternative model. This assumption is tested using the cumulative Webhost-repertoire data from each student-user. The Webhost-trajectory is more likely to be curvilinear than the Website-trajectory so that if this fails to be curvilinear then the Website-trajectory will also fail to be curvilinear (since Website-repertoire $\geqslant$ Webhost-repertoire).

Figure 3.4 illustrates the *raw* Website-trajectory and Webhost-trajectory for a student-user during study-year two (selected as an illustration because the study-year click rate of 1,179 clicks is close to the average of 1,173 clicks). The raw trajectory function is only defined once each session, not click by click as can be achieved with instrumentation and is given directly by the first stage analysis of the Web log by cumulating the session click rate (page 73). Given a minimum of two sessions than a best fit straight line can be approximated to the trajectories. The fit (using gnuplot (Williams & Kelley, 1998)) has not been constrained so the graphs illustrate non-zero intercepts.

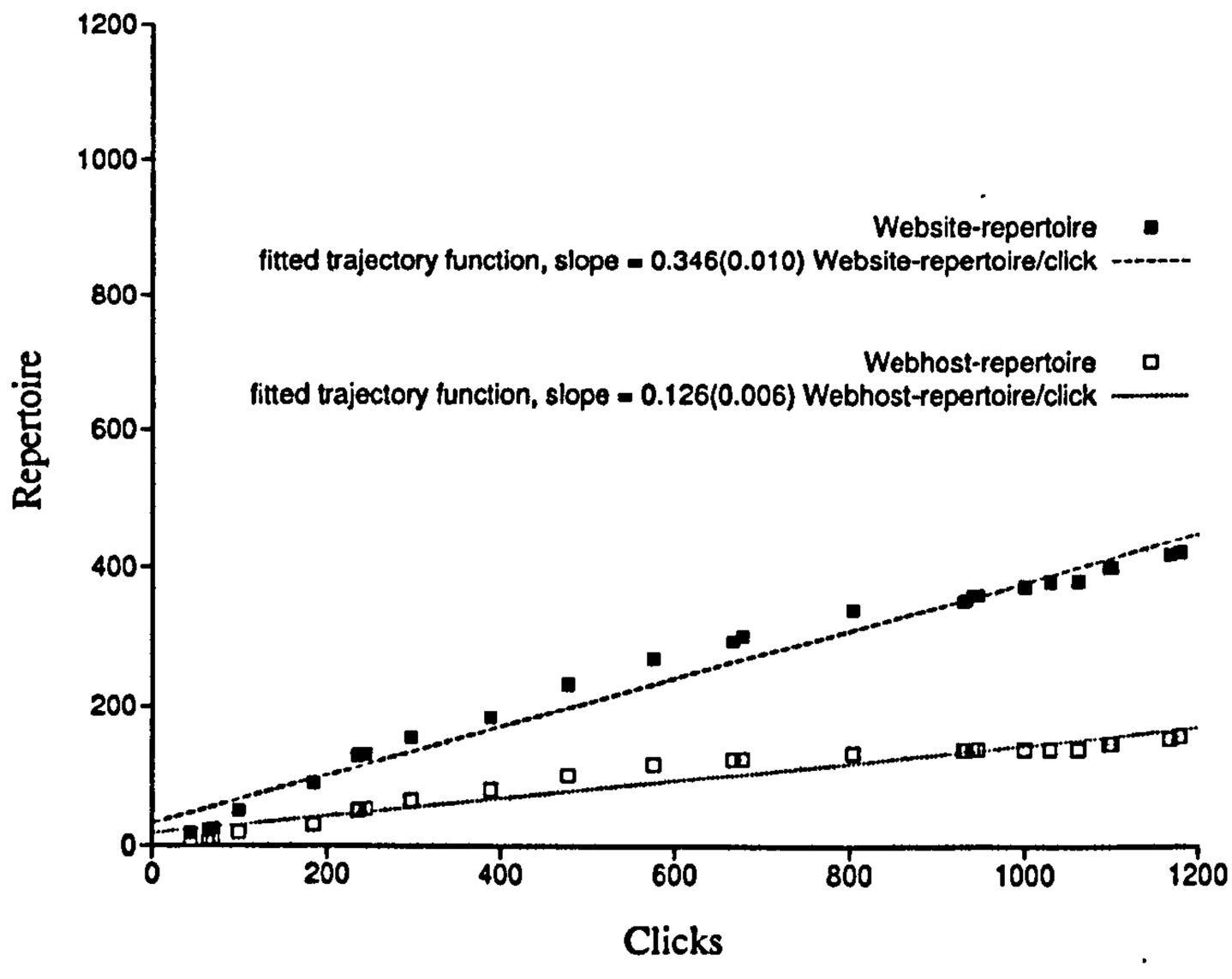

Figure 3.4: Example of Website and Webhost trajectories

If the random walk model (or Heaps' Law) is an improvement in describing repertoire development compared to using just the raw trajectory as in Figure 3.4, then, on

average the straight lines fitted to the transformed data will be closer to the observed data than in the case of the raw data. Figure 3.5 shows the straight lines fitted to the logarithmically transformed Website-trajectory and Webhost-trajectory for the same student-user as in Figure 3.4

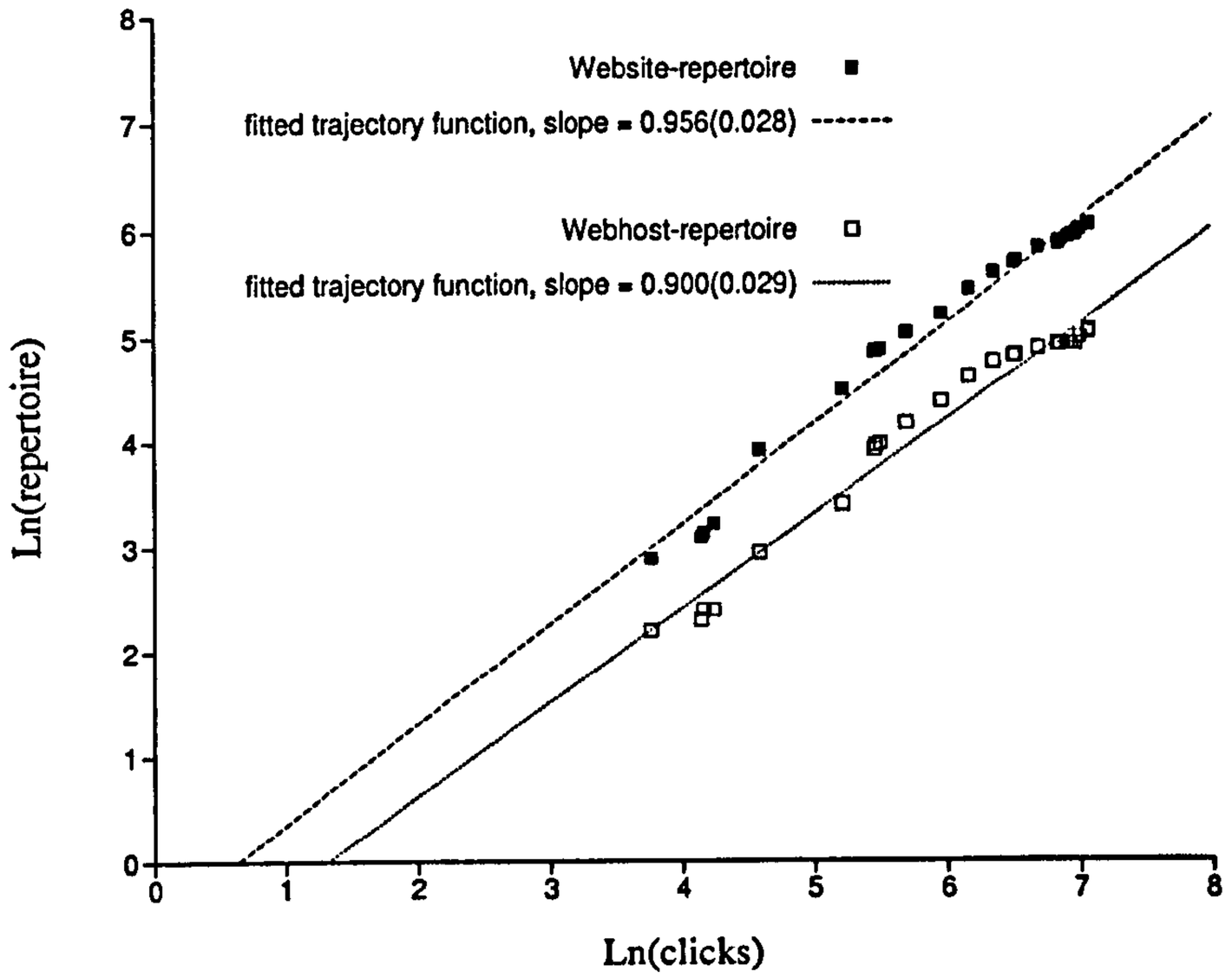

Figure 3.5: Example of logarithmically transformed Website and Webhost trajectories

How close the straight line is to the data is measured by the standard error proportion of the slope. Excluding cases where the data has only two data points (and hence there is no error), the mean standard error proportions for the raw Webhost-trajectories for all the student-users during study-years one and two are 6.8%(0.3%) and 5.6%(0.3%) respectively (the sample size exceeds 1,000 for each study-year). When logarithmically transformed these means are 6.4%(0.2%) and 5.3%(0.2%). Although these are numerically smaller, the z test of difference is not significant ($\alpha = 0.05$) during either study-year ($z_{\text{study-year one}} = 1.1$, $z_{\text{study-year two}} = 0.83$) hence it is accepted that the logarithmic transform procedure does not give an improvement in characterising repertoire growth.

The trajectory data of this investigation are therefore modelled linearly by,

$$\tau(n) \propto n, \text{ or,}$$

$$\tau(n) = Tn + constant \text{ where } 0 \leqslant T \leqslant 1.$$

Even in the more favourable circumstances of the Webhosts there is no evidence to prefer a curvilinear development of repertoire. This could be because the recurrence effect is small or because there is a large threshold value for cumulative clicks before recurrence becomes effective. Figures 3.6 and 3.7 illustrate the Website and Webhost trajectories for student-users during study-year two.[21] The graphs confirm that there is no generally appearing downward curvature for large cumulative clicks as is predicted by the random walk model. The few instances of downward curvature that do appear (in Figure 3.7) confirm the rationale for seeking empirical confirmation of recurrence generally by reference to the Webhost-trajectories, and the absence of recurrence generally. Although it is evident that recurrent trajectories do arise from Web information seeking, identifying the conditions under which they occur is beyond the scope of this investigation.

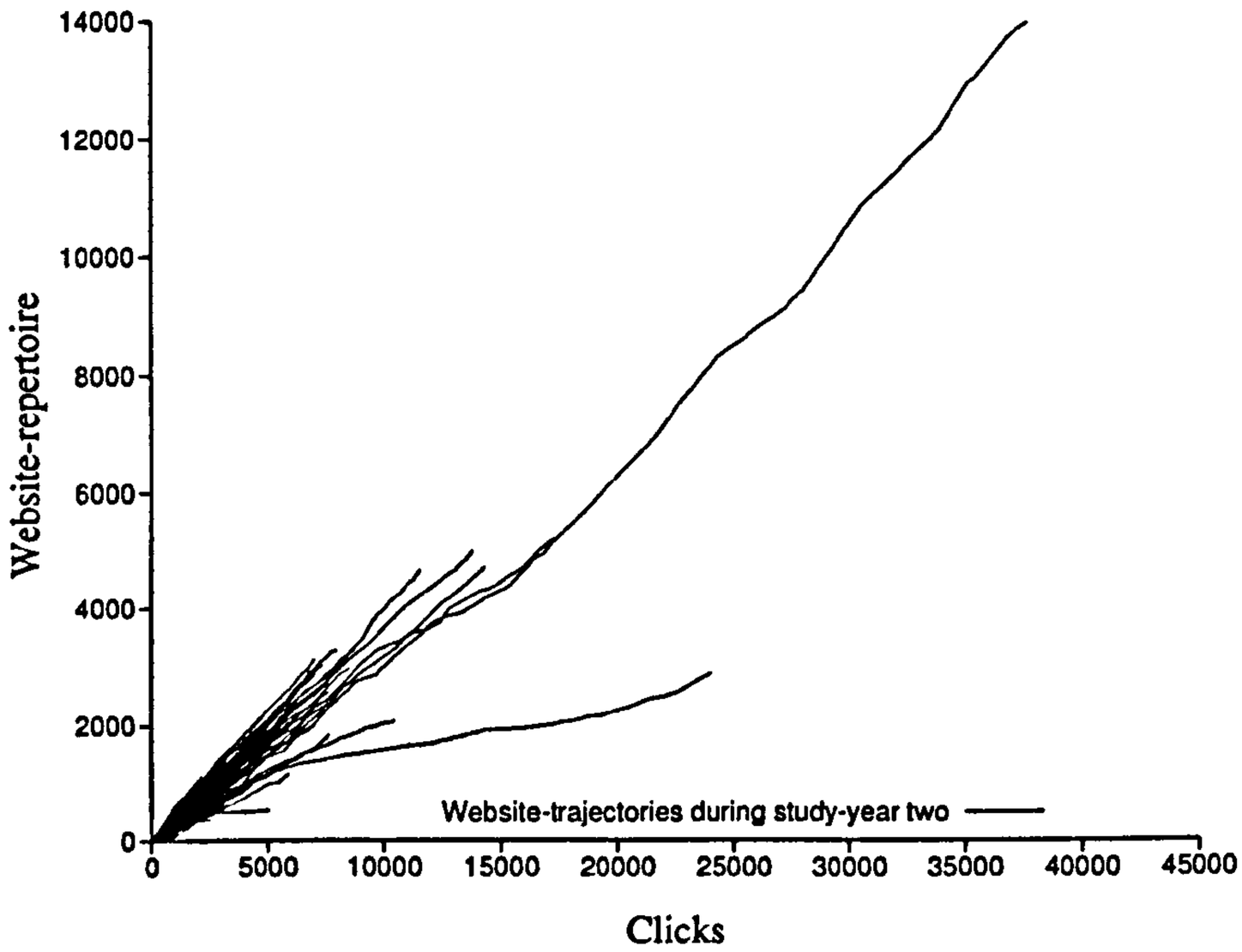

Figure 3.6: Overlay of Website-trajectories during study-year two

[21] The equivalent graphs in respect of study-year one are illustrated in Appendix B.

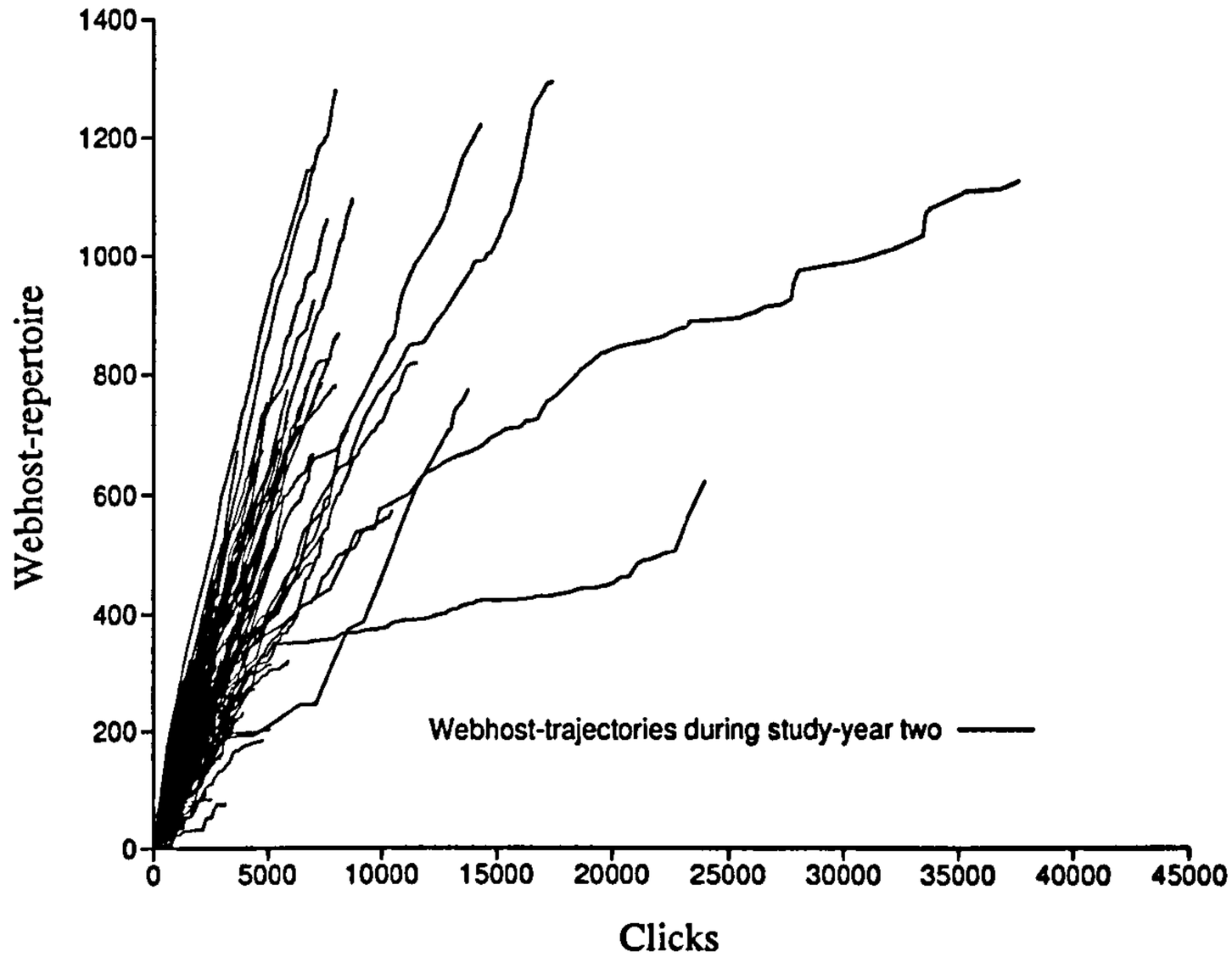

Figure 3.7: Overlay of Webhost-trajectories during study-year two

Tauscher (1996, pp. 51–52) uses the metrics,

$$R = \frac{\text{clicks} - \text{repertoire}}{\text{clicks}}, \text{ and,}$$

$$C = 1 - R$$

which she refers to as the recurrence and composition rates.

Hence Tauscher's C, for Websites, would be found as $\frac{\text{study-year Website-repertoire}}{\sum \text{session click rate}}$ from the second stage Web log dataset described on page 73. Thus C is based on only the last datum point of the observed trajectory while the metric T which is used here to characterize how student-users seek Web information is based uniformly on the Web information seeking activity during each of a student-user's sessions.

3.5 Web session-conformance

The idea of the Web session-conformance extends the interpretation of the Web transaction log by evaluating the *similarity/dissimilarity* of each session. Since each session is a bag of Websites complete with associated frequencies of occurrence then this suggests using a vector model approach. In this application the usual focus

on retrieving documents for users is inverted in order to describe users by their documents (for example, Frants, Kamenoff & Shapiro, 1993; Fu *et al.*, 1999).

The goal of the session-conformance metric is to evaluate each session by comparing it with the average session. If Web information seeking sessions in respect of the frequencies of visits to Websites during the session were all the same then the session-conformance metric should be the same for all sessions. On the other hand if each session were different, then each session-conformance metric should be different. The session-conformance metric should also enable student-users to be compared both within and between the study-years. This goal is achieved using a tf*idf procedure as follows.

In principle each of the 46,558 sessions during both study-years can be represented as a position vector, or point, in a Website space hyper-quadrant where the co-ordinates of the point are the frequencies of occurrence of the Websites in the session. (Co-incident points represent sessions which are the same.) The complete Website space thus has a dimension equal to 507,618 which is the overall Website repertoire of the study.

The quadrant of two dimensional Website space showing twelve sessions is illustrated in Figure 3.8. The twelve session vectors are given in Table 3.5 and include three co-incident sessions, A1, B1 and D3. In addition the position vectors of two groups of sessions are co-linear, for example C1 and C2 since $\vec{c_2} = (0,2) = 2(0,1) = 2.\vec{c_1}$

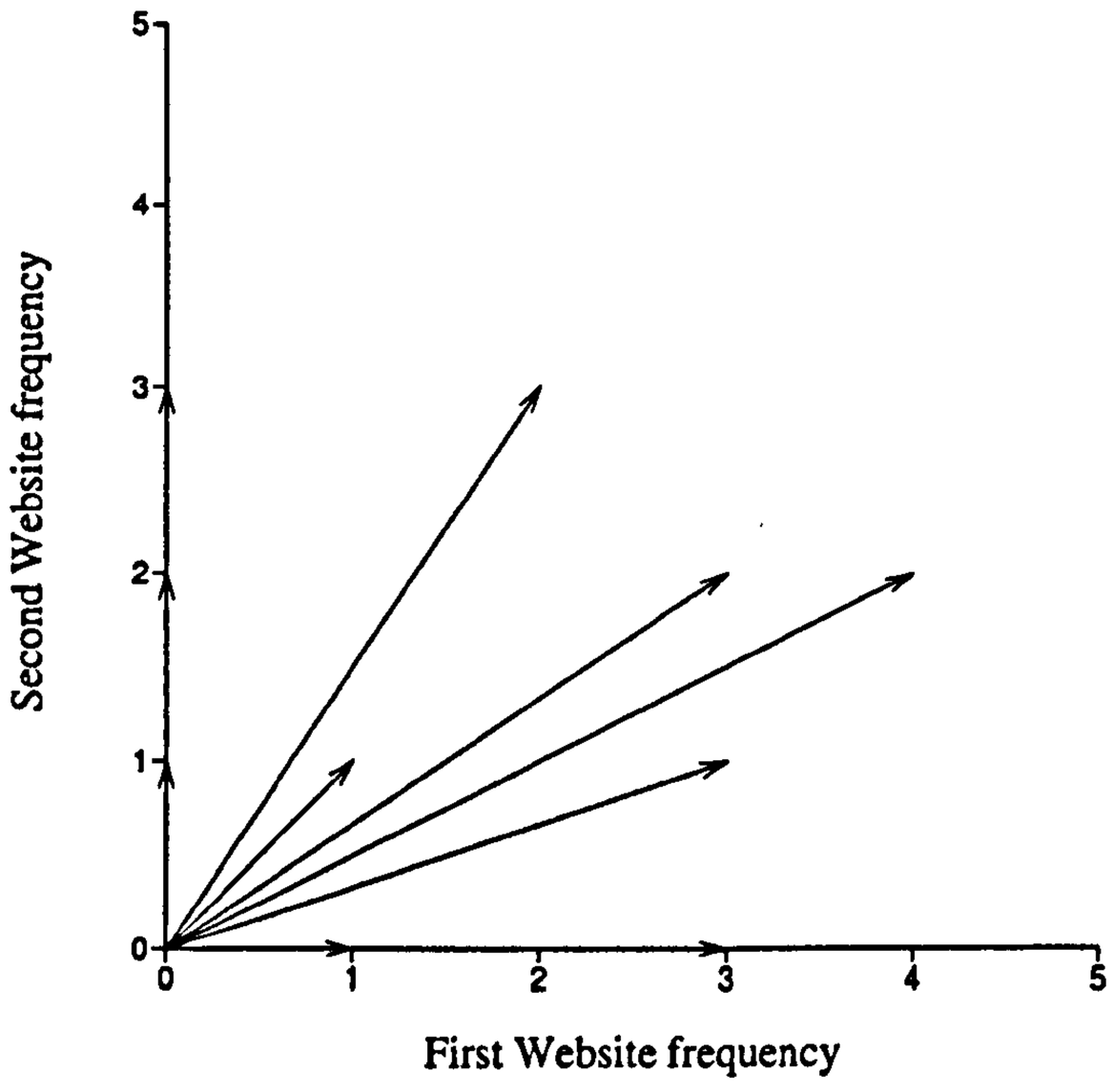

Figure 3.8: Session position vectors in a quadrant of two dimensional Website space

Session	Position vector Website frequencies	
	First Website	Second Website
A1	1	0
A2	3	2
B1	1	0
B2	4	2
C1	0	1
C2	0	2
C3	1	1
C4	0	3
C5	3	0
D1	2	3
D2	3	1
D3	1	0
Total	**19**	**15**

Table 3.5: Two dimensional Website position vectors

The size of the complete Website space is reduced to 1,001 dimensions by ranking all the Websites by their session frequency (as is required by the tf*idf procedure) and considering only those Websites ranked in the top one thousand. Figure 3.9 shows the Zipf distribution of the Website session frequency which has the expected power law appearance. The top ranked (=1) Website occurs in 7,896 sessions.

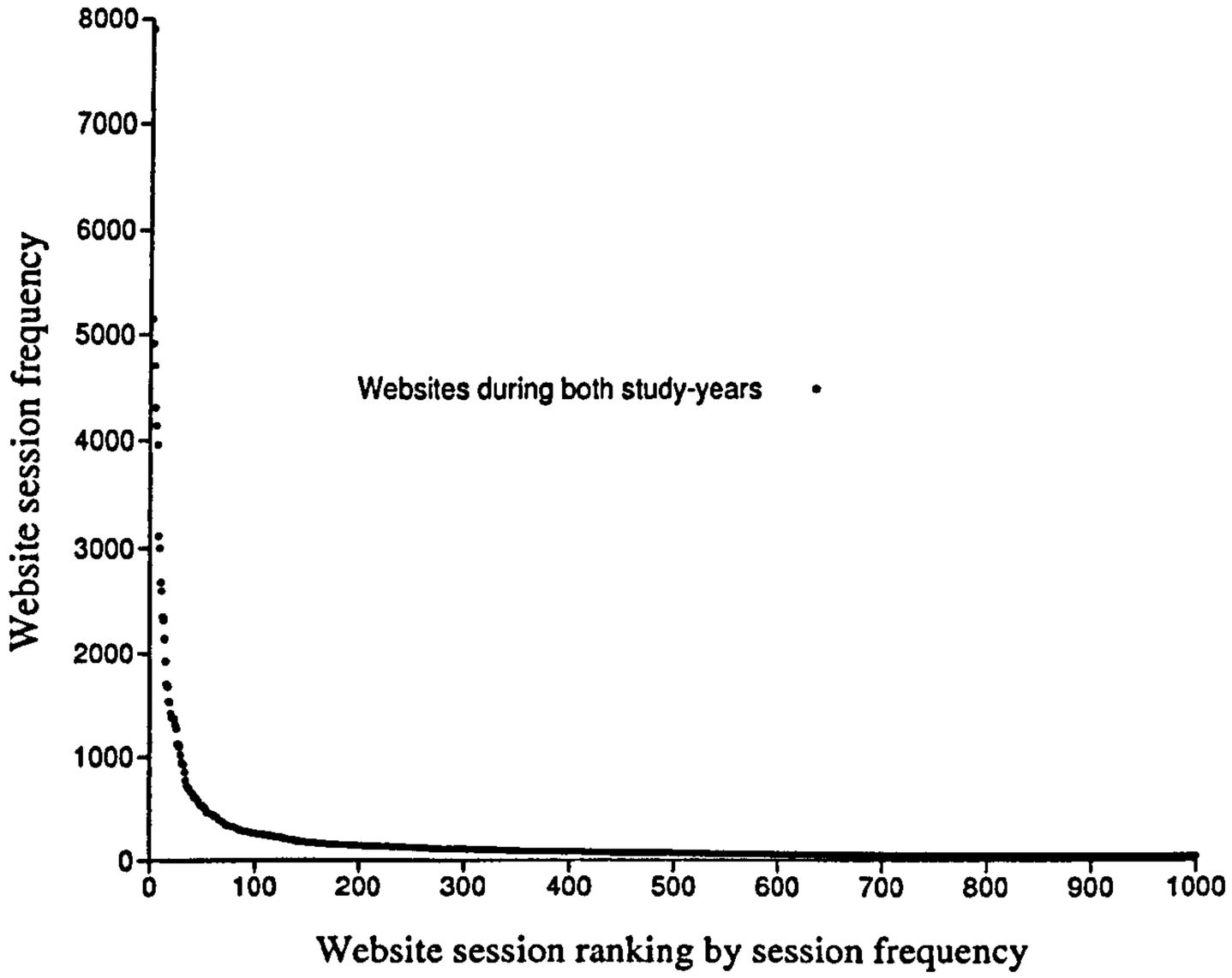

Figure 3.9: Zipf distribution of the top one thousand ranked Websites

Websites ranked greater than 1,000 each occur in less than 41 sessions but collectively these 506,618 (= 507,618 − 1,000) Websites have a session frequency of 41,161 sessions and 423,752 Websites (83% of the overall Website repertoire) at rank 83,866 are visited during only one session. Individually therefore these 506,618 Websites are *rarely* visited and here are referred to as *rare* Websites. However 2,674 (=13%) sessions during study-year one and 3,605 (=14%) sessions during study-year two comprise Web requests just to rarely visited Websites.

Hence Websites ranked > 1,000 cannot be ignored. In respect of session-conformance all of these Websites are regarded as a single Website with session frequency 41,161.

Each session is thus represented by a position vector ($tf_1, tf_2, \ldots tf_{1,001}$) where tf_i is the frequency of occurrence of the ith ranked Website in the session. A tf*idf weighted position vector (Manning & Schütze, 1999) for each session is given by,

$$((1 + log(tf_1)).log(\frac{N}{df_1}).tf_1, \ldots (1 + log(tf_{1,001})).log(\frac{N}{df_{1,001}}).tf_{1,001})$$

where df_i is the session frequency of the ith ranked Website and $N = 46,558$ (that is the overall session frequency).

When each weighted position vector is normalized to be of unit length, the normalized weighted position vectors correspond to points distributed on the surface of a

hyperquadrant. This is illustrated by the normalized weighted position vectors of the previous two dimensional example which are illustrated in Figure 3.10.

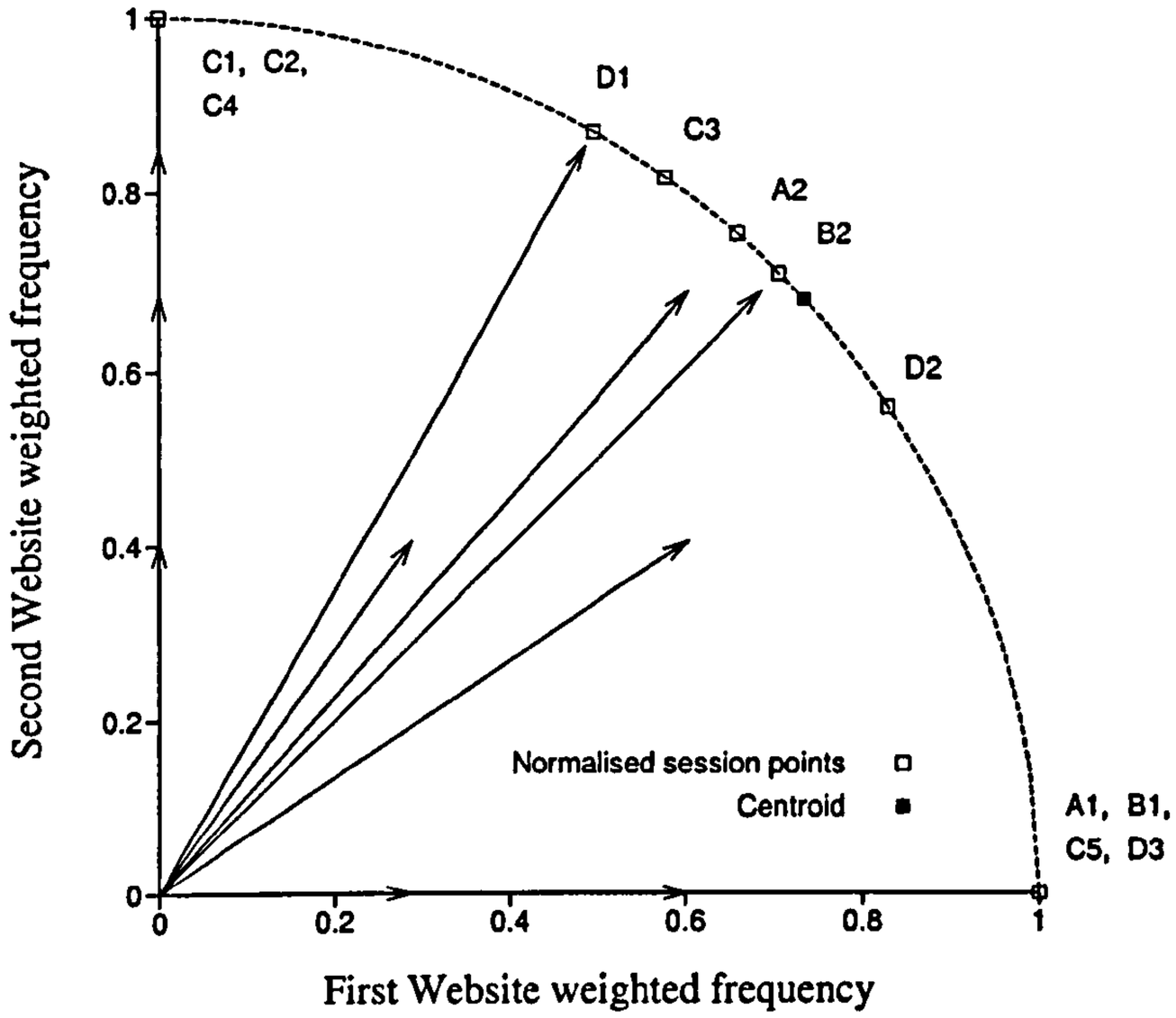

Figure 3.10: tf*idf weighted session vectors and their normal projections in a quadrant of two dimensional Website space

Table 3.6 shows the tf*idf weighted position vectors derived from Table 3.5. The Table also shows the normalized weighted position vectors. For example, for session A2, the tf*idf vector is given by,

$$(3,2) \mapsto ((1+log(3)).log(\frac{12}{9}).3,\ (1+log(2)).log(\frac{12}{8}).2)$$

$$= (0.603733132,\ 0.686512105)$$

which has length = 0.914216914. The normalized weighted position vector for A2 is thus,

$$(0.660382808,\ 0.750929122) = \frac{1}{0.914216914}(0.603733132,\ 0.686512105).$$

Table 3.6 also gives the normalised weighted vectors which shows that sessions which have co-incident position vectors have co-incident normalised weighted vectors. Sessions which have co-linear position vectors also have co-incident normalised weighted vectors.

Session	tf*idf vector (1+log(tf)).log(N/df)		Normalised tf*idf vector	
A1	0.287682072	0	1	0
A2	0.603733132	0.686512105	0.660382808	0.750929122
B1	0.287682072	0	1	0
B2	0.686494107	0.686512105	0.707097512	0.70711605
C1	0	0.405465108	0	1
C2	0	0.686512105	0	1
C3	0.287682072	0.405465108	0.578656895	0.815571087
C4	0	0.850914059	0	1
C5	0.603733132	0	1	0
D1	0.487088090	0.850914059	0.496793476	0.867868793
D2	0.603733132	0.405465108	0.830156577	0.557530320
D3	0.287682072	0	1	0

Table 3.6: Normalized weighted two dimensional Website position vectors

The centroid of the normalized weighted position vectors is found as $\frac{1}{46,558}(\sum w_1, \ldots \sum w_{1,001})$ where w_i is the ith component of each normalised weighted position vector, or, for the twelve sessions in the two dimensional example, ($\frac{7.273087268}{12}, \frac{6.699015372}{12}$) which is re-normalized as (0.735539098, 0.677482277). The centroid for the example is illustrated in Figure 3.10 above.

The squared Euclidean displacement between the (normalised) centroid and each normalised weighted position evaluates or measures each session by comparing it with the average session. Table 3.7 gives these measures for the twelve sessions in the two dimensional example.

	Squared centroid-session displacement
A1	0.528921804
A2	0.011042907
B1	0.528921804
B2	0.001687084
C1	0.645035446
C2	0.645035446
C3	0.043680545
C4	0.645035446
C5	0.528921804
D1	0.093246498
D2	0.023340939
D3	0.528921804

Table 3.7: Conformance computation example

It is clear that, as in Table 3.7, pairs of sessions which have co-linear position vectors in the underlying Website space will have the same squared centroid-session displacements. However it is not clear that two different position vectors in the underlying Website space will have different squared centroid-session displacements (which is needed in order to meet the goal of the session-conformance metric). In practice it is found that the 21,366 sessions from study-year one comprise 16,045 different normalised weighted position vectors which generate 15,990 unique displacements. For study-year two there are 25,192 sessions out of which there are 19,134 different normalised weighted position vectors and 19,082 unique displacements. Inspection of the 107 instances (55 during study-year one and 52 during study-year two) when two apparently different normalised weighted position vectors have the same displacement measure shows that the differences between the normalised weighted position vectors are just an arithmetic artifact (in the fifteenth decimal place of a term frequency) which sometimes arises when co-linear position vectors are normalised. Therefore it is concluded that, during study-year one there are 15,990 different sessions (75%) out of 21,366 and during study-year one there are 19,082 different sessions (76%) out of 25,192 each having a unique squared centroid-session displacement.

Hence the squared centroid-session displacement satisfies the goal for a session-conformance metric set out earlier. Comparisons can be made both within and between study-years since the centroid reference position is the average for all 46,558 during both study-years. The distribution of squared centroid-session displacements has a minimum of 0.442135360738888 and ranges up to 1.99600236415299. For convenience this is adjusted[22] to have a minimum of zero.

[22] Arithmetic operations on document similarity measures derived by using the vector model

The Web session-conformance metric for each session is thus defined to be this adjusted squared centroid-session displacement. Hence the *larger* the metric the further apart or more *dissimilar* are the sessions in respect of the Websites visited.

The centroid estimates the average session and therefore consists of a portion of each of the 1,001 Website used in the vector model computation. A consequence of this is that sessions comprised exclusively of rare Websites (that is ranked 1,001) are closest to the average session (since at 41,161, they have the largest *df*) and therefore have a session-conformance metric of zero. The next most conforming sessions with a session-conformance = 0.478288221731976 are sessions which comprise just the most frequently occurring Website (rank = 1) and a rare Website at rank 1,001.

The session which is closest to (most resembles) the mean session-conformance (= 1.1400927729225) consists of visits to the Websites ranked 1, 4, 14, 24, 54, 117, 267, 372 and 1,001.

The most distant session, that is the session with the largest session-conformance metric (= 1.55386700341411) which least resembles the centroid or average session consists of a visit just to the Website at rank 883.

Sessions which include visits to a rarely visited Website together with one or more of the most frequently visited Websites are thus closer to the centroid or average session (and have a small session-conformance metric) while sessions which comprise only visits to the less frequently visited (but not rarely visited) Websites are more distant from the average session (and have a large session-conformance metric).

Since each Web information seeking session is associated with a particular student-user than the mean and range of the session-conformance metrics for that student-user can be found. In the two dimensional case, for example the mean (un-adjusted) metric is 0.269982356 in respect of A.

Cothey (2002) uses a procedure similar to that described here to measure the Web-host conformance of Web information seeking sessions which adopts a Boolean approach but uses a binary weighting scheme allied to ranking rather than tf*idf (Aalbersberg, 1994).

It is argued earlier that Website requests to rare Websites cannot be ignored. These requests occur in 41,161 out of 46,558 sessions overall. As described above these

procedure are problematic because the measures are not *transitive* and the procedure is directionless, that is several non-identical documents may be equidistant from a given reference point and hence appear to be equally similar (Zhang & Korfhage, 1999). The uniqueness property of the Web conformance computation overcomes this problem so that the Web session-conformance is a bijection from the weighted normalised space to the closed interval [0,2] in the real line which can therefore be relocated.

sessions dominate the session-conformance computation and show that sessions typically (but not necessarily) comprise visits to *rare* Websites in conjunction with none or more visits to the more frequently visited Websites. A small (non-zero) session-conformance corresponds to a session containing visits to highly ranked Websites; a larger session-conformance corresponds to visiting less highly ranked Websites. (A zero session-conformance corresponds to visiting just rare Websites.) This prompts the question of what is the effect of adopting an alternative approach. In the first instance the criterion of rank > 1,000 might be varied, either up or down. In the second instance rare Website visits might be ignored.

The boundary criterion of rank > 1,000 to qualify as rare is arbitrary but its choice needs to recognise the computational precision which is available when calculating the session-conformance. Clearly if the criterion is weakened, say rare $_{500}$ is rank > 500, then more sessions will comprise just visits to rare $_{500}$ Websites and therefore will have zero session-conformance. Also the discrimination of the session-conformance (as regards between more frequently by-session visited Websites) is reduced. Under these circumstances the session-conformance metric is mostly distinguishing between different combinations of 500 rather than 1,000 Websites visited. Later (see Chapter four) student-users are partitioned into those whose sessions *always* include visits to the more frequently by-session visited Websites and those student-users who sometimes have sessions which do not include any visits to any of these more frequently visited Websites (that is, have a session with zero session-conformance). Weakening the boundary criterion would push student-users from the former partition into the latter. Since it is found that student-users tend to migrate in this direction then weakening the boundary criterion would tend to inflate this finding.

If the criterion is strengthened, say rare $_{2000}$ is rank > 2,000, then the opposite effect occurs (assuming a sufficient computational precision). Fewer sessions will contain visits to rare $_{2000}$ Websites so these will dominate less[23] the computation and the session-conformance will mostly distinguish combinations of 2,000 Websites visited. The number of student-users who sometimes have sessions which do not include any visits to the 2,000 Websites most frequently visited by-session would reduce (compared with rare $_{1000}$). Although the relative sizes of the partitions mentioned would therefore change the migration of student-users would remain in evidence. This is because there are no constraints on a student-user having a Web information seeking session which comprises just visits to Websites arbitrarily low down in the ranking of the entire Website vocabulary.

[23] Since as the criterion is strengthened the session frequency of rare Websites reduces then there may be some large value at which the session frequency < 7,896 which is the session frequency of the top ranked Website. If this occurs then rare Websites would no longer dominate the session-conformance computation. However it is seen that in the strongest instance there are 423,752 Websites at rank 83,866 which each occur in a single session. In order for the session frequency < 7,896 an infeasible average > 54 of these Web requests is needed within each such session.

If the visiting of rare Websites were to be ignored then where would the boundary be drawn and what would be the consequence? For rare $_{1000}$ the session frequency of the rare Websites is 41,161 and about 14% of sessions during each of the study-years comprise just visits to these rare Websites. If the session-conformance computation ignored these Web requests then the metric would distinguish about 86% of sessions by their content of the 1,000 more highly ranked Websites and it would now be affected most strongly by the top ranked Website which has a session frequency of 7,896. A zero session-conformance would now correspond to visiting just this Website. As before small session-conformances would correspond to sessions containing visits to high ranked Websites with larger (more dissimilar) session-conformances corresponding to sessions which comprise visits to one or more low ranked Websites. Also as before the most dissimilar session would be just a visit to a single low ranked Website. Moving the boundary criterion in these circumstances clearly has the effect of changing the proportion of sessions which are ignored but does not otherwise change the description given. Thus if the 14% of ignored sessions were uniformly distributed among all student-users then discarding them may not be material to the investigation. However it is found that although when considered by-session Web information seeking typically consists of visiting rare Websites and none or more highly ranked Websites, when considered by-user this is no longer the case. In particular some, but not all, student-users never have sessions which contain only visits to rare Websites. These student-users would no longer be distinguishable were these sessions to be ignored.

This investigation focuses attention on these 14% of sessions and especially how their distribution among student-users changes between during study-years one and two.

3.6 Longitudinal-developmental analysis

The design of many longitudinal-developmental investigations, particularly in education, is based on administering an assessment procedure at regular intervals throughout the study. Such longitudinal data can therefore be treated as a function of time. In this investigation, as is illustrated by Figure 3.2 (see page 64), individual student-users' Web information seeking is irregular. Some student-users have many sessions during a study-year but most have few and the intervals between sessions vary so that the primary data from the Web log analysis is unbalanced (the number of data points is not the same for all the student-users) and is not related to time. However the balanced meta-analytic data for each individual represents the student-user's information seeking *during* each of the two study-years. These data can therefore be analysed *conditionally* (Goldstein, 1979), that is a value for during study-year

two can be considered as a function of a value for during study-year one (which is therefore fixed, hence the conditionality). This analytic procedure contrasts with considering both values to be a function of time.

The regression of study-year two on study-year one provides an estimate of the functional relationship and in particular is a technique for analyzing the phenomenon of change over time (Plewis, 1985). The linear regression can be found as a least squares best fit straight line. The simple regression of study-year two on study-year one is biased in that it considers the two study-years asymmetrically and privileges change in the forward direction. The simple regression of study-year one on study-year two is similarly biased in the reverse direction. In order to overcome the bias which in the forward direction tends to deflate the measure of change (that is the gradient of the regression) then a symmetrical least squares procedure is used. This symmetrical *conditional-regression* is a least squares best fit on a sum and difference transformation of the original independent and dependent variables. Following this least squares best fit analysis then the transformation is inverted so that the value for study-year two is expressed as a function of the value for study-year one.

The form of longitudinal-development analysis which is used here is a symmetrical conditional-regression of a single Web information seeking metric on itself which is here referred to as a *conditional analysis*. Hence if for each student-user there were no change in a particular Web information seeking metric, M, between during study-year one and during study-year two, then $M_{\text{study-year two}} = M_{\text{study-year one}}$ and the slope of the conditional regression line would be one.

Figure 3.11 shows the conditional distribution of student-user's session rate metric (the number of sessions during a study-year) where the session rate during study-year two is considered as a function of the session rate during study-year one. The two straight lines shown are the *no-change line*, study-year two = study-year one, and the conditional-regression of study-year two on study-year one. The session rate metric is not considered a reliable[24] indicator of Web information seeking activity but is used only to provide an example to illustrate the conditional analysis used here for longitudinal-developmental analysis. This is similar to the longitudinal-developmental analysis in Cothey (2002) however the earlier analysis does not use the symmetrical regression which is employed here.

[24] There will be external factors not investigated here which affect whether or not a student-user undertakes a Web-information seeking session; the investigation here is concerned only with activity when a session is undertaken.

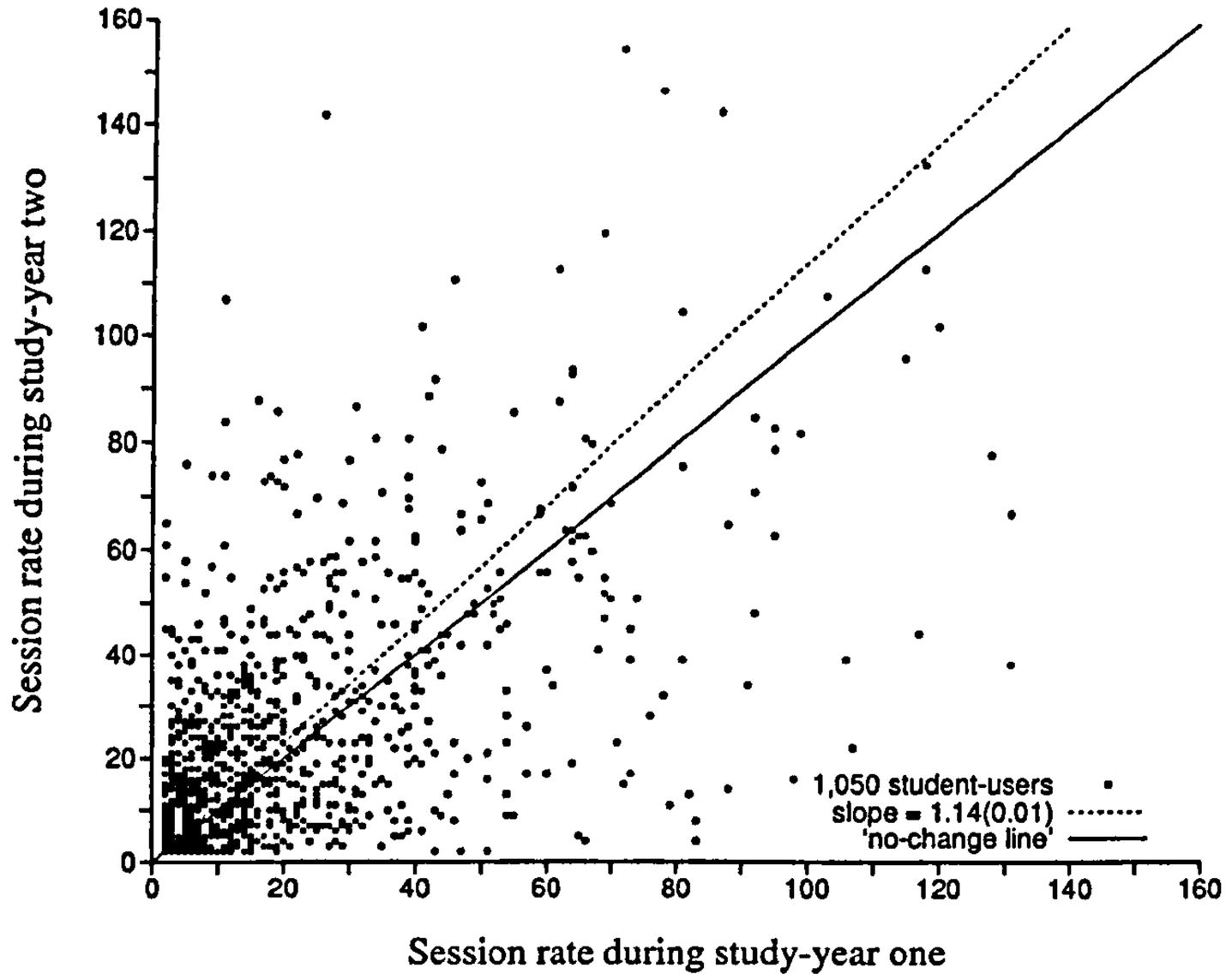

Figure 3.11: Conditional distribution of student-user's session rate

The slope of the conditional-regression illustrated in Figure 3.11 is 1.14(0.01). Thus, when giving equal weight to each of the 1,050 student-users, the best (linear) estimate of the proportional change in a student-user's session rate is that the rate during study-year two is 1.14 × the student-user's session rate during study-year one. The standard error of the slope is 0.01 so that the regression slope is significantly more than one ($p < .00,\ z = 14$). Thus the null hypothesis of no change is rejected and the alternative, that the regression slope is more than one, is accepted. Hence it is concluded that student-users' individually *increase* their session rate during study-year two compared with during study-year one.

The mean values for the session rate are 20.35(0.65) sessions and 23.99(0.69) sessions for study-years one and two respectively and the mean paired difference is 3.64(0.63) sessions which also suggests that session rate has increased. This nearly 18% increase exceeds the 14% increase indicated by the conditional analysis but this can be explained because the smaller differences associated with student-users who have small session rates are overwhelmed by the larger differences associated with student-users who have greater session rates.

It has already been mentioned that the characterizations of student-users' Web information seeking activity are distorted by the skewed nature of the frequency distribution of the characterization metrics. Most student-users, during most sessions do not *energetically* seek Web information, that is the session click rate is generally

small as illustrated in Figure 3.3 (see page 64). The mean values of characterization metrics generally overstate how individual student-users locate Web information.

The conditional-regression technique gives equal weight to each student-user regardless of how energetic that student-user is, and evaluates change as being a proportional effect rather than a difference effect. Hence for this study conditional analysis is more appropriate than a paired difference analysis for investigating the longitudinal-developmental change in how student-users locate Web information.

Conditional-analyses can also compare the change which applies in different groups of student-users. For example, the change effect in men student-users can be compared with the change effect in women student-users as illustrated by Figure 3.12 which shows the previous conditional distribution of session rates partitioned by student-user gender. This reveals differences in the longitudinal-development of men and women student-users.

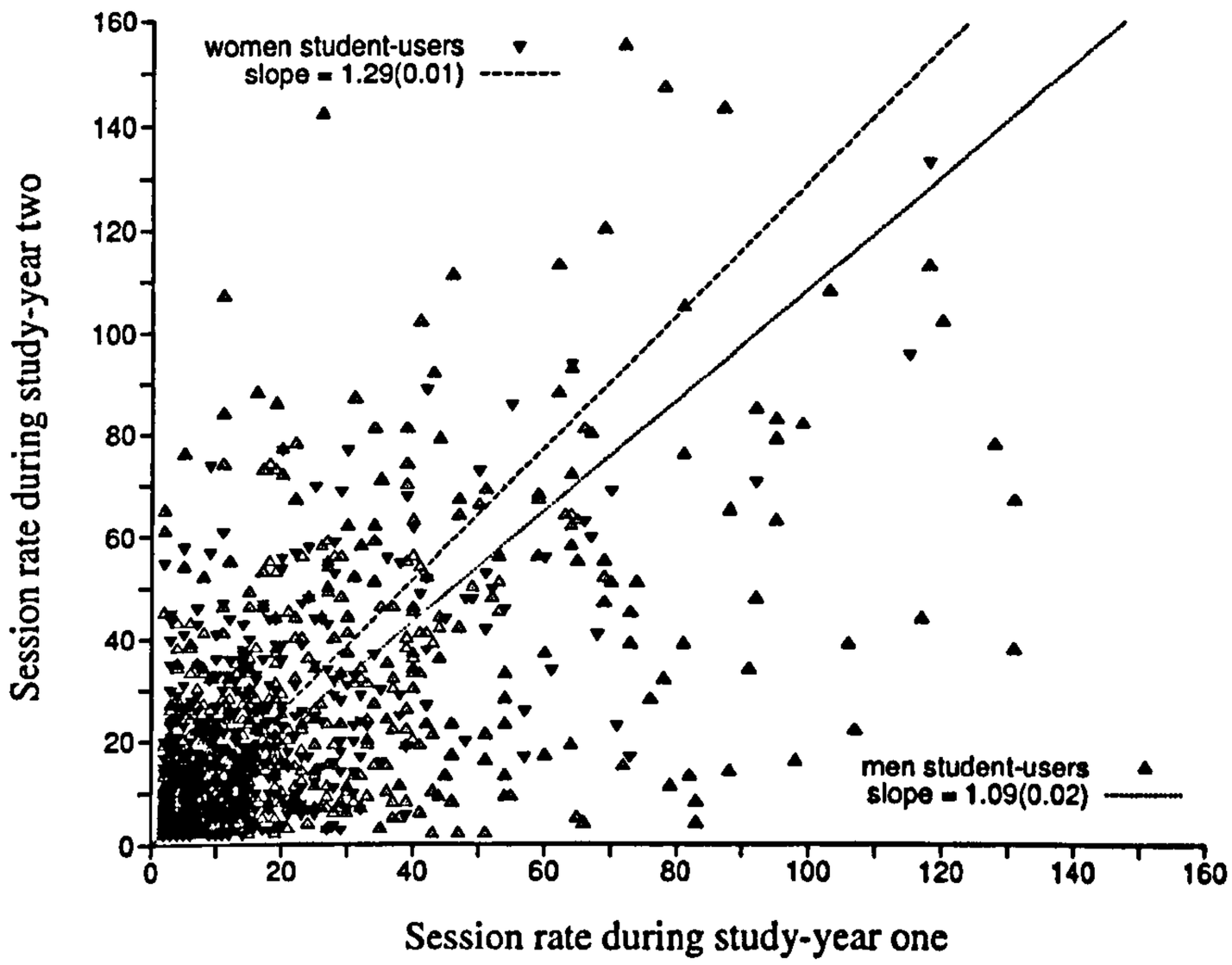

Figure 3.12: Conditional distributions of student-user's session rate by-gender

The conditional-regression slope for the women is 1.29(0.01) which is significantly different from one ($p < .00,\ z = 29$) so it is accepted that they increase their session rates. However the conditional-regression slope for men is only 1.09(0.02), thus although men student-users also significantly increase their session rate ($p < .01,\ z = 4.5$) they do so to a lesser extent than women. The conditional analysis shows that the longitudinal development of men and women student-users' session rate is significantly different ($p < .01,\ z = 8.9$).

The conditional-regression technique can therefore both demonstrate the existence of a longitudinal-developmental change effect and can compare such effects as they apply to different groups of student-users. In particular the technique is used to explore a *novice-effect* where the rate of change for novices is an exaggeration of the rate of change for more experienced users, see Chapter six.

3.7 Summary and discussion

The method for discovering how student-users locate Web information operationalises *student-users locating Web information* as *Website* visiting activity. This activity is determined from client-side Web logs collected unobtrusively from 1,050 full-time undergraduate student-users over a period of two successive academic years or study-years (1998/1999 and 1999/2000). Hence the method facilitates both a longitudinal analysis and a repeat study whereby the reliability of results is improved by them being reproduced. The data collection method exploits the internal data capture mechanism of the browser history file. This client-side Web log overcomes a common difficulty with server-side based Web research which is an inability to capture users' Web information requests that are satisfied by their local browser cache. Anonymous student-user identifying codes allow separate individual analysis of each student-user's Web log of Website visiting. A Website is defined by *conditioning* the url-strings which student-users submit to request Web information so that requests for the same resource under a different url label, for example `<url:http://www.bris.ac.uk/>` and `<url:http://Bris.ac.uk/>` can be identified. A *Webhost* is the unique DNS official name of of the Website server.

Web information is defined to exclude the use of Web based email systems, chat-rooms, the institution's Web OPAC and the institution's own Web based teaching and learning material. The results of the Web log analysis are therefore not distorted by, for example, how much local Web based material student-users request as part of their academic courses. Conditioning also excludes *embedded* files, typically image files or advertising banners, which are not expressly requested and makes adjustments for requesting internal-anchors within Web pages.

The Web log analysis is based on *daily* sessions of Web information seeking activity. The 1,050 student-users are all the (full-time undergraduate) student-users in the 1997 and 1998 registration cohorts who recorded at least two daily sessions of Web information seeking activity during each of the two study-years.

Each student-user's activity is analysed both session-by-session and session-to-session. Hence both isolated and extended information behaviour is considered. The session-by-session analyses are based on metrics such as the *average Webhost-persistence*

(measured in Websites per Webhost) which is defined as;

$$\frac{\sum \text{session Website-repertoire}}{\sum \text{session Webhost-repertoire}}.$$

The session Website-repertoire and session Webhost-repertoire are the number of different Websites and Webhosts respectively which a student-user visits during a (daily) session. Since the ratio $\frac{\text{session Website-repertoire}}{\text{session Webhost-repertoire}}$ for each session gives the average number of different Websites visited within each different Webhost (during that session) then the notion of Webhost-persistence is similar to that of *path length* found in the literature (see Chapter two).

Session-to-session analyses use metrics based on the student-user's Website *trajectory* (or Website repertoire growth curve) and *session-conformance*. These metrics take account of the student-user's extended information seeking activity over multiple sessions. For example, the slope of the Website trajectory distinguishes student-users who do a lot of Website revisiting overall (in respect of their own Website vocabulary) from those who do not. The session-conformance metric, which uses the vector model to analyse sessions based on the Websites which are visited, considers the overall similarity/dissimilarity of a particular session in the context of all the sessions (by all of the student-users). Hence the variety of Websites which each student-user visits during a session can be compared both within that student's collection of sessions and among sessions generally.

A feature of Web log data generally which is also found here is that it is distorted (when compared with a typical Gaussian distribution). In consequence nonparametric statistical techniques are used here to test conjectures relating to differences in student-user's Web information seeking activity metrics both between groups of student-users and over time. The metrics which are used are chosen to allow similarities, differences and changes in student-users' Web information seeking to be interpreted so that a narrative profile is developed which describes how student-users locate Web information.

Interpretation of the Web log is made difficult by the structural change which may be occurring in the Web. *Conditional analysis* is used to examine the phenomenon of change by modelling metric values for study-year two as a function of their value during study-year one. Comparison of the change for different groups of student-users can help to disambiguate some of the effects of structural change and non-structural change since a structural phenomenon should apply to all users.

The method for discovering how student-users locate Web information validly and reliably determines on a large scale what it is that the student-users do. The analyses are user-focussed and extended. They thus meet the research needs identified

in Chapter two. The findings from the analyses are expressed in the form of narrative descriptions of how student-users locate Web information and are set within a framework of a personal Web information infrastructure.

4

How do student-users use the Web?

4.1 Introduction

Some overall dimensions of the Web log have been mentioned previously. It comprises all the 1,990,488 requests for Web information from 1,050 student-users during two academic years (or study-years). The Web log is organised into daily Web information seeking sessions of which there are 46,558. Hence, typically,

> each student-user undertakes twenty-two sessions of Web information seeking per study-year, and,
>
> during each session a student-user makes forty-three clicks, or requests forty-three items of Web information.

However the goal of this investigation is to provide a fuller interpretation of the Web log than this, and to reveal something of the variety of how student-users locate Web information. In particular the analytic strategy of the investigation considers the changes, similarities and differences in how groups of student-users locate Web information both between the groups, among the groups, and over time. The analyses are based on *user-characterization* metrics and *user-attributes*. The user-characterization metrics each measure a particular feature of locating Web information in respect of an individual student-user. For example, the *average session click rate*, say 65.5 clicks per session, estimates for an individual student-user how many requests for Web information he (or she) makes during each session.

The user-attributes which apply here are *gender* together with *session-rate* (*smaller/larger*) and *conformance* (*conformant/eclectic*) which are discussed below. Each user-attribute generates a dichotomous partitioning of the set of 1,050 student-users, for example

the gender partitions are 540 men and 510 women student-users. In Chapter six a *novice* attribute is introduced in the context of the longitudinal-developmental investigation.

The conformance based user-characterizations and user attribute are founded on a method for interpreting the Web log which draws on the vector model used in information retrieval. The way in which the vector model is used is discussed in Chapter three.

Change over time in how student-users use the Web may be because of change in (a) Web information seeking tasks, (b) the *structure* of the Web including its information seeking affordances and session duration or (c) the individual.

Task and individual differences are discussed in Chapter three and individual change, in particular a *novice-effect* is examined in Chapter six. Differences or change in information seeking user-characterizations which apply across all the user-attribute groups of student-users may be interpreted as a structural effect. On the other hand, change and differences which affect some but not all of the user-attribute groups are less likely to have a structural origin.

This Chapter has three principal Sections which are;

User-characterization in which the user-characterizations are defined and used to analyse the Web log in order to illuminate different aspects of how student-users locate Web information. This Section provides some initial answers to the question of how student-users locate Web information and reveals considerable variation. The analysis also discovers that student-users fall naturally into two distinct conformance groups, those whose Web information seeking always includes visits to the more frequently visited Websites and those who are more eclectic in how they locate Web information.

Similarities and differences in which groups of student-users defined by their user-attributes are compared in order to discover whether or not they are the same in how they locate Web information. This comparison also facilitates some disambiguation of the effects of structural change, and,

Website popularity in which the investigation of the extent to which different groups of student-users share the Websites which they visit during each study-year is reported. This is the least meta-analytic part of the investigation and identifies particular Websites so that an increase in diversity in how student-users locate Web information is revealed.

A Summary and discussion describing how student-users use the Web to locate information concludes the Chapter.

4.2 User-characterization

There are seven user-characterization metrics which describe and differentiate how student-users locate Web information. These are,

1. average session click rate,
2. average query-click proportion,
3. average Website-re-request rate,
4. average Webhost-persistence,
5. Website-trajectory slope,
6. average session-conformance, and
7. session-conformance range.

Each of these is now defined and used to provide an initial overall description of how student-users locate Web information. In the next Section these user-characterization metrics are employed in conjunction with the user-attributes to examine potential similarities and differences between different student-user groups in how they locate Web information.

Five of the seven user-characterization metric are *session-by-session* metrics, that is, each of a student-user's sessions is analysed separately. The Website-trajectory slope and the session-conformance range user-characterization metrics are *session-to-session* metrics in that they describe and differentiate how student-users locate Web information by analysing the cumulative effects of their Web information seeking activity. For example, the Website-trajectory slope analyses the extent to which a student-user visits Websites which are previously unvisited by him (or her) during an entire study-year.

Average session click rate

The number of clicks, that is Web requests, made by each student-user during each session is reliably captured by the client-side Web log which includes requests satisfied by the local browser cache. The issue of caching is discussed in Chapters two and three. The *click rate* metric is the number of clicks during a given time period, and in particular the *session click rate* metric, which is measured in clicks per session, is the number of clicks during each session. The power law frequency distribution

of the session click rate metric and the consequential distortions in how student-users locate Web information is discussed in Chapter three. The session click rate frequency distribution is illustrated previously, see Figure 3.3.

Since the (anonymous) student-user who originates each session is known, Web information seeking sessions can be grouped *by-user*. Hence, for each student-user, the user-characterization, *average session click rate* is computed as,

$$\text{average session click rate} = \frac{\sum \text{session click rate}}{\text{session rate}}, \text{ (clicks per session)}$$

$$= \frac{1}{n} \sum_{\text{sessions}\,=\,1}^{n} \text{session click rate, (clicks per session).}$$

The average session click rate user-characterization for a student-user is found for each study-year by computing the summation of the sessions over one study-year only. For example a student-user who, during a study-year, undertakes five information seeking sessions which comprise eight, four, twelve, nine and thirteen clicks respectively has an average session click rate of $\frac{46}{5} = 9.2$ clicks per session.

During study-year one the average session click rate for student-users (to the nearest click) ranges from 2 to 511 clicks per session and has a mean of 29.0(0.85) clicks per session. During study-year two the range is 2 to 598 clicks per session and the mean is 44.1(1.11) clicks per session. The increase is significant ($p < .001,\ z = 10.8$).

The session click rate metric does not distinguish between Websites so this increase may reflect additional requests to the same Websites or even a reduction in the number of different Websites being visited. Student-users are more *energetic* that is submit more clicks per session during study-year two compared with study-year one but this may be because the duration of sessions is longer or because the Web is structurally different in study-year two and several clicks are needed where a single click sufficed during study-year one. (Only explicit Web requests by student-users are included in the Web log analysis; implicit requests such as for *embedded* images are filtered out, see Chapter three. Therefore, for example, more graphical content in Web page design would not explain the increase in click rate.)

The frequency distributions of the average session click rate user-characterization for each study-year are illustrated in Figure 4.1. The median values are about 21 and 35 clicks per session during study-year one and two respectively. Thus most student-users submit fewer Web requests during a session than does the mean student-user.

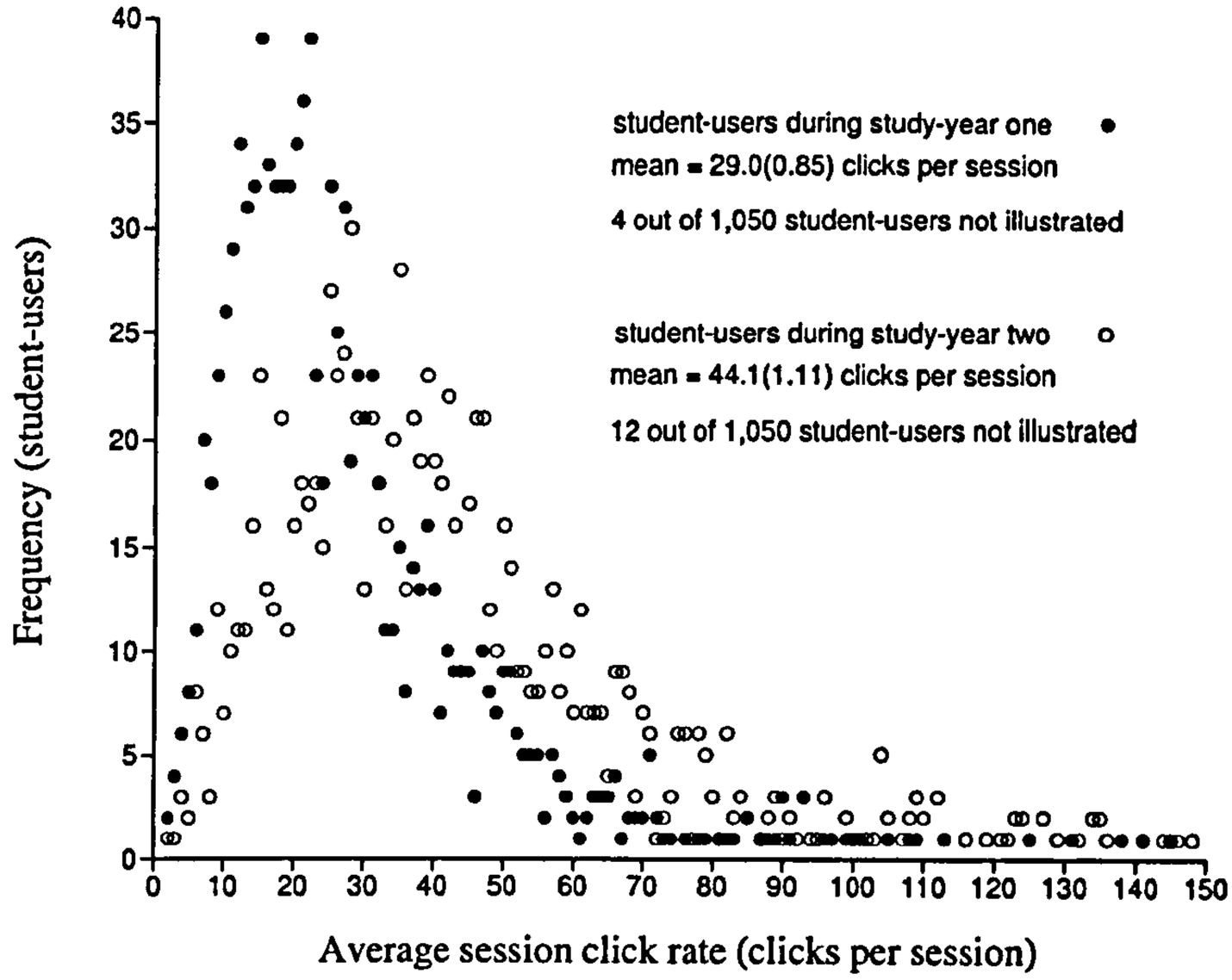

Figure 4.1: Frequency distributions of student-user's average session click rate (range illustrated up to 150 clicks per session)

It is found that those student-users who have larger session rates are also more energetic. This is discussed later.

Average query-click proportion

A *query-click* is a Web request (click) which includes a *search-part.* Search parts are delimited by the "?" character, or query, so that for example, the url-string, `<http://www.foo.com/path/search?bar>` contains the search part consisting of `<bar>`. *Querying*, or submitting query-clicks, is taken as indicating a more active form of Web information seeking by a student-user than passively link clicking, see Chapter three

The average query-click proportion user-characterization, is focussed on the *occurrence* of a search-part. Some particular Websites to which search-parts are submitted together with the content of search-parts is examined in Chapter five where how student-users use Web information location services is investigated.

The *query-click rate* metric is the number of query-clicks during a given period. The *session query-click rate* metric is the number of query-clicks during each session. This is measured in query-clicks per session. Figure 4.2 illustrates the frequency distribution of the session query-click rate metric. The distribution resembles that

of the session click rate metric which is illustrated previously in Chapter three (see Figure 3.3). This is inevitable because the session query-click rate is constrained above by the session click rate (since session query-click rate $\not>$ session click rate).

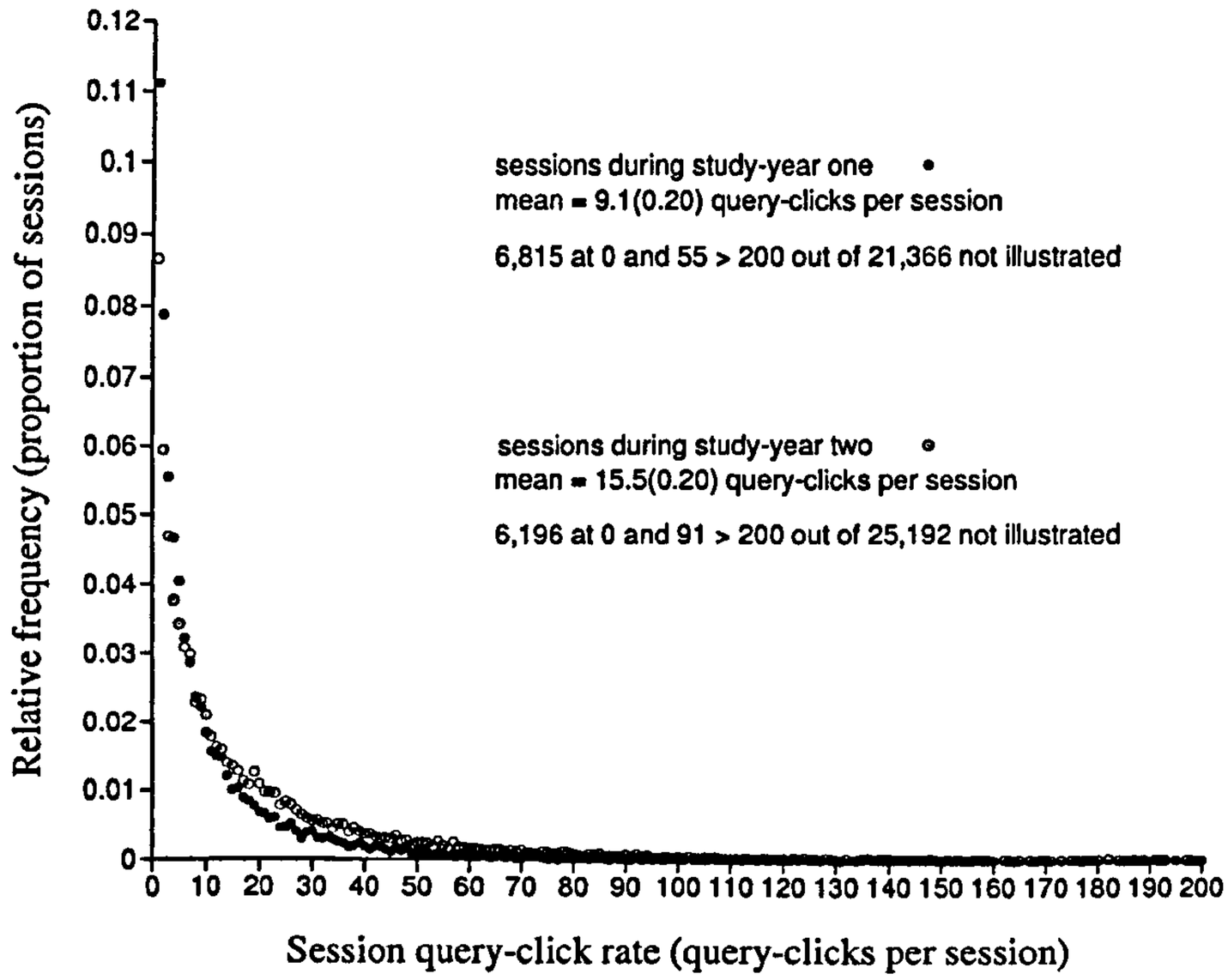

Figure 4.2: Frequency distributions of session query-click rate (range illustrated up to 200 query-clicks per session)

During each study-year more than about a quarter of all sessions involve no querying and during each study-year there are instances of very high session query-click rates (>1000 query-clicks per session). These unusual sessions arise when a student-user is apparently carrying out a bibliographic search of a particular online database which generates a multiplicity of individually different Websites rather than the more common search-parts attached to a single Website. In consequence these differences in the url-strings are not *conditioned* to be equivalent, see Chapter three.

The increase in the mean session query-click rate between during study-year one, 9.1(0.20) query-clicks per session, and during study-year two, 15.5(0.20) query-clicks per session, is significant ($p < .001$, $z = 22.6$). Manifestly, student-users during study-year two are doing a lot more querying but more analysis is needed in order to offer an interpretation.

The *query-click proportion* metric, measured in query-clicks per click, is the proportion of clicks which are query-clicks. For each student-user, the user-characterization

average query-click proportion is computed as,

$$\text{average query-click proportion} = \frac{\sum \text{session query-click rate}}{\sum \text{session click rate}}, \text{ (query-clicks per click).}$$

Like the average session click rate user-characterization, the average query-click proportion user-characterization for a student-user is found for each study-year by computing the summation of the sessions over one study-year only. The frequency distributions of student-users' average query-click proportion during each of study-years one and two are illustrated in Figure 4.3. These have the expected[1] Gaussian form. The mean value during study-year one is 0.26(0.003) query-clicks per click which increases to 0.33(0.004) query-clicks per click during study-year two. The increase is significant ($p < .001$, $z = 14.0$) and is evident as a sideways shift in the graph. In a few (five) instances student-users never use query-clicks and stand out in the graph as having a zero average query-click proportion.

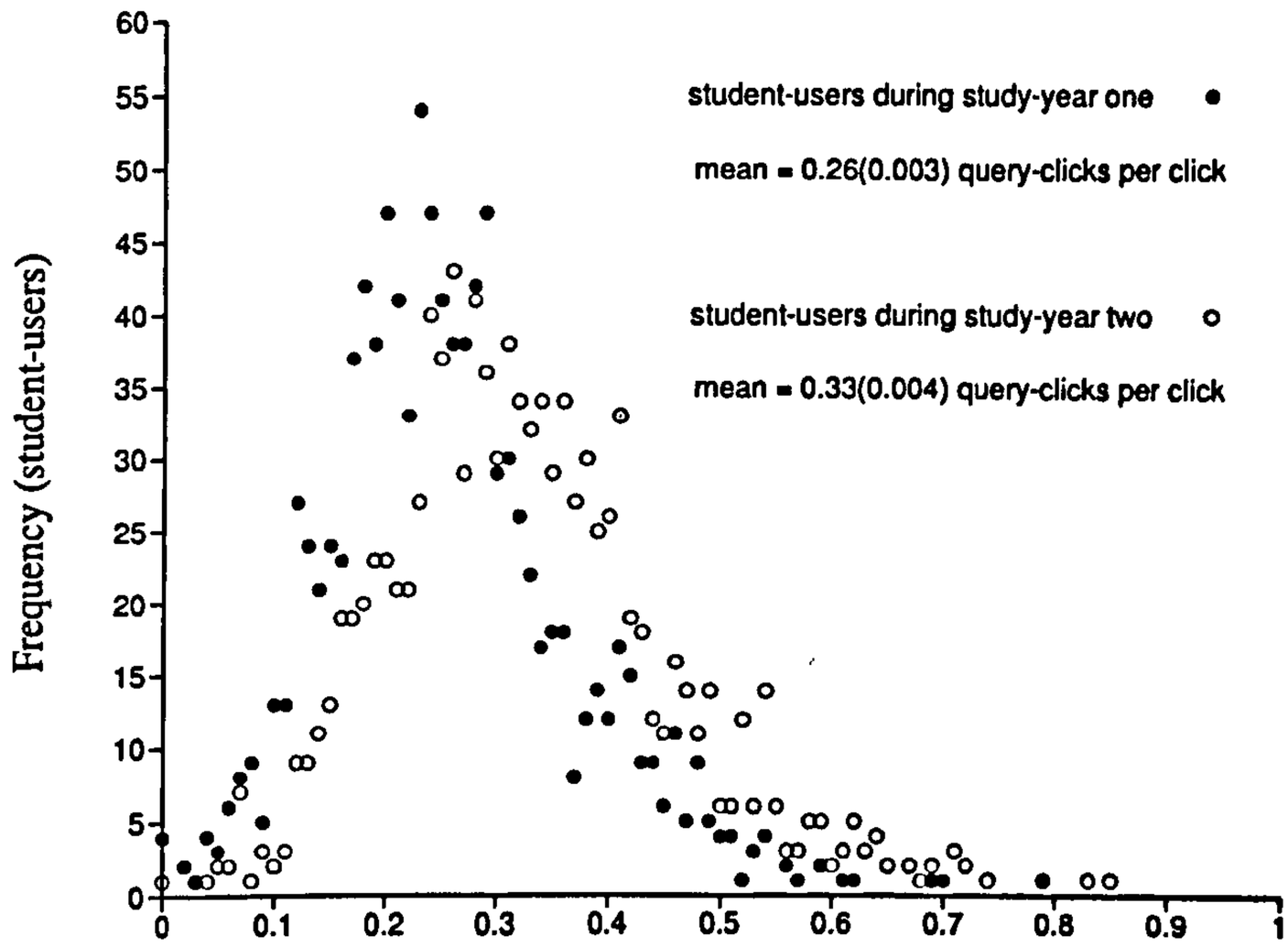

Figure 4.3: Frequency distributions of student-user's average query-click proportion

On average, student-users during study-year two are therefore more *active* in how they locate Web information, that is, their Web information seeking involves more querying (rather than just passive link clicking). How this querying is distributed among sessions is investigated with the *query-session rate* metric. A *query-session* is a session which includes a query-click so, for example, the count of the number

[1] From the central limit theorem since the distribution is effectively sample means.

of query-sessions by a student-user during a study-year is the student-user's query-session rate measured in query-sessions per study-year. The frequency distributions of the query-session rate, which are analogous to the session rate distributions discussed in Chapter three (see Figure 3.2) are illustrated in Figure C.1 in Appendix C.

The *query-session proportion* metric, measured in query-sessions per session, for each student-user is the ratio of the student-user's query-session rate to session rate during a particular study-year. The frequency distributions of student-user's query-session proportions (rounded to the nearest 5%) are illustrated in Figure 4.4. The mean value of 0.70(0.007) query-sessions per session during study-year one increases significantly ($p < .001,\ z = 7.6$) to 0.77(0.006) query-sessions per session during study-year two. The graph shows the many student-users (about 170) during each study-year who have a query-session proportion of one, that is they use querying during each of their Web information seeking sessions and the few who, since they never use query-clicks, have a zero query-session proportion.

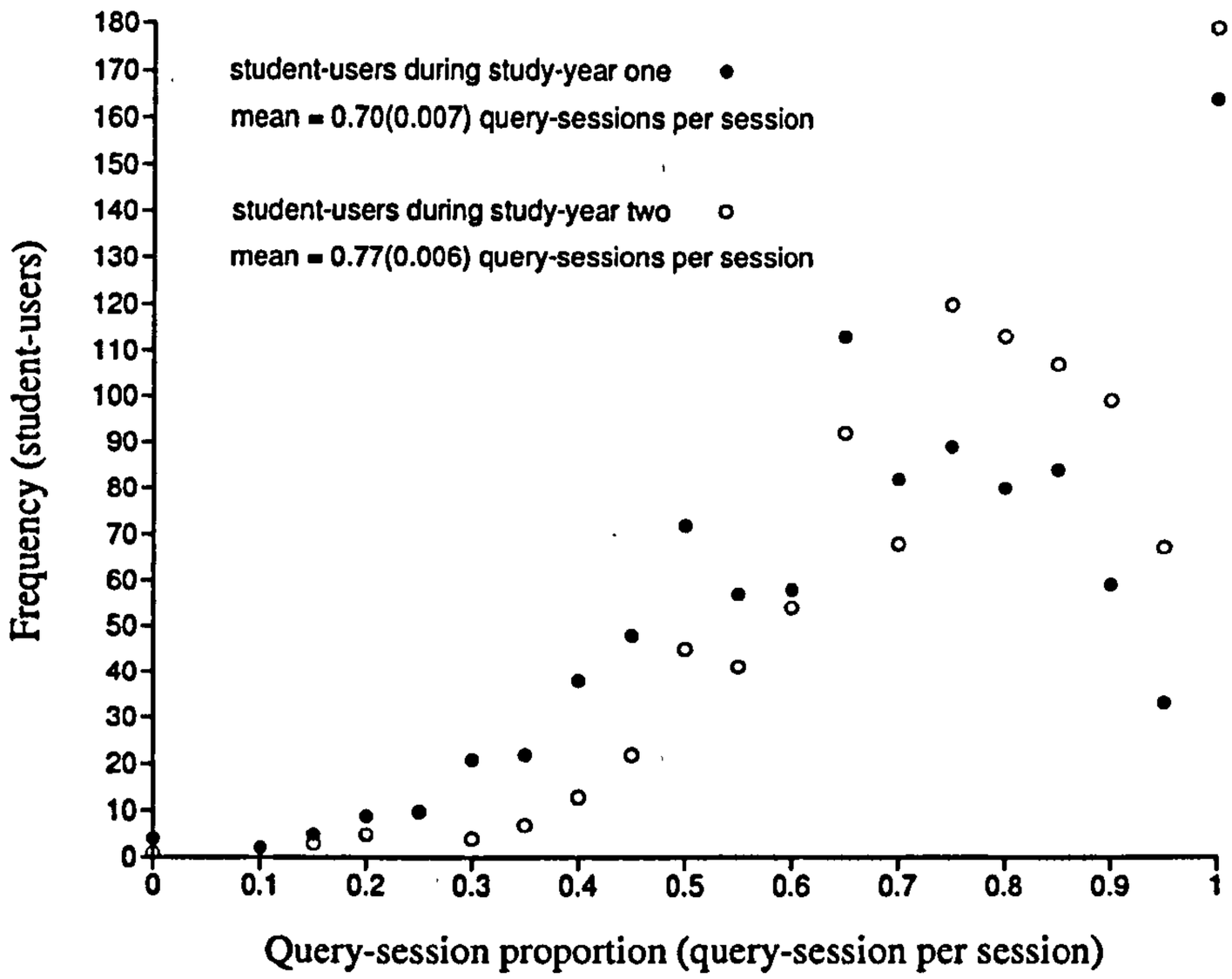

Figure 4.4: Frequency distributions of student-user's query-session proportion

The distribution of the query-session proportion frequencies also shows that the number of student-users with small query-session proportions reduces between during study-year one and during study-year two. However there still remain too many sessions which are not query-sessions if one supposes that query-clicks are uniformly distributed across all sessions. Conservatively, taking the probability that a click is a query-click to be 0.2 and the session click rate to be only ten clicks per session, then

the probability that a session is a query-session is about 0.9. As either the query-click probability or session click rate increases then this 0.9 probability increases. Since the mean query-session proportion < 0.8 then it appears that a bias is operating and that query-clicks are being concentrated in some query-sessions so as to reduce the overall query-session proportion.

During study-year one the overall query-click proportion is 26% ($= \frac{194,232}{758,636}$) query-clicks per click which increases to 32% ($= \frac{389,136}{1,231,852}$) query-clicks per click during study-year two. The overall query-session proportions are 68% ($= \frac{14,551}{21,366}$) query-sessions per session and 75% ($= \frac{18,996}{25,192}$) query-sessions per session during study-years one and two respectively (see Tables C.1 and C.2 in Appendix C).

Hence querying, whether indicated by query-clicks or query-sessions, by student-users generally increased between during study-years one and two. But the distribution of query-clicks across sessions is not uniform and, for example, querying is later discovered to be more pronounced among the groups of conformant and the smaller session rate student-users.

Average Website-re-request rate

The *average Website-re-request rate* and *average Webhost-persistence* user-characterizations both make use of analyses of the conditioned url-strings in the Web log. Each Web request refers to a *Website* within a *Webhost*. The Website is defined by the conditioned url-string which excludes any search-part. For example, the conditioned url-string, `<foo.com/path/search?bar>` comprises the Website `<foo.com/path/search>` and the (conditioned) Webhost `<foo.com>`. The conditioning of url-strings which standardises the representation of Websites and Webhosts for analysis is discussed in Chapter three.

The *Website-re-request rate* metric, measured in clicks per Website, is the ratio of the number of Web requests (or clicks) to the number of different Websites visited (that is, the *Website-repertoire*). Hence the metric indicates how much individual Websites are revisited. During a given period, in particular a session, a student-user's session Website-re-request rate is,

$$\text{session Website-re-request rate} = \frac{\text{session click rate}}{\text{session Website-repertoire}}, \text{ (clicks per Website).}$$

The *average*[2] *Website-re-request rate* user-characterization of a student-user during

[2] This average is not the arithmetic mean but can be interpreted as an average gradient.

a study-year is computed as,

$$\text{average Website-re-request rate} = \frac{\sum \text{session click rate}}{\sum \text{session Website-repertoire}}, \text{ (clicks per Website)}$$

where the summation is over all the student-user's sessions during the study-year.

The frequency distributions for the average Website-re-request rate for student-users during each of study-years one and two are illustrated in Figure 4.5. The mean average Website-re-request rate increases significantly ($p < .001$, $z = 13.4$) from 2.0(0.01) to 2.3(0.02) clicks per Website between during study-years one and two. However this 15% increase in the metric is not as large as the 50% increase in average session click rate which means that the more energetic Web information seeking of student-users during study-year two is not explained just by an increase in revisits to Websites, although this does explain some of the increase.

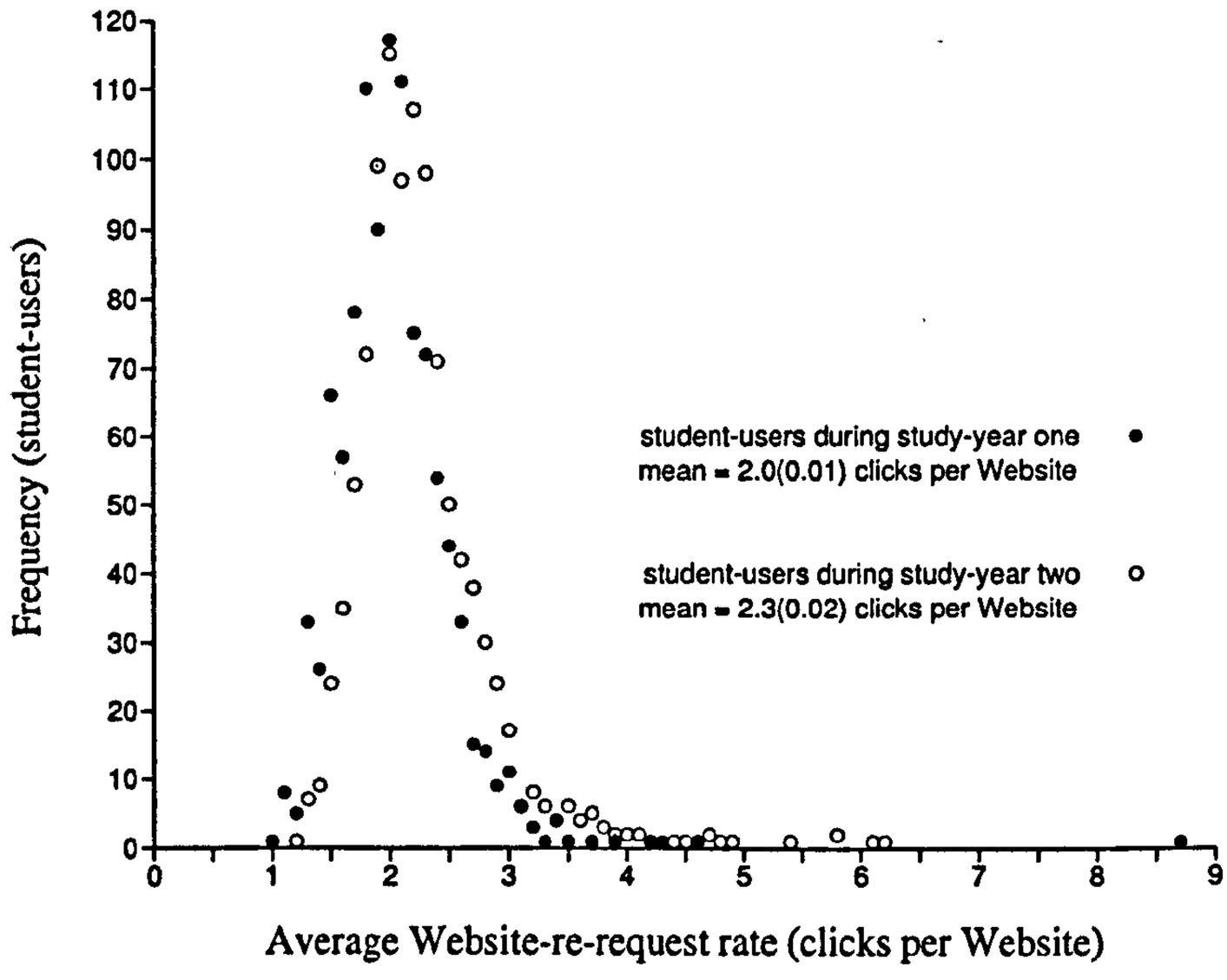

Figure 4.5: Frequency distributions of student-user's average Website-re-request rate

It is also possible that the Website re-request rate may be influenced by change in Web affordances and by structural change.[3] The graph in Figure 4.5 shows that student-users vary greatly in their Website re-requesting rate. It is found later for example, that although there is little association generally, for student-users with

[3] Web page designers could introduce more query-click based interaction which would increase the Website-re-request rate metric. Cache-busting techniques such as page timeouts would not affect the metric since requests satisfied by the local browser cache are included, see Chapters two and three.

small average session click rate, average Website-re-request rate may be associated with average session click rate.

Website re-requesting indicates something about how a student-user is locating Web information with respect to particular sessions. Thus it differentiates those student-users who request many Web pages (for example alternating back and forth between the same two Websites) from those who request many *different* Web pages during a session. The *Website-trajectory slope* user-characterization, which is discussed below, is needed to distinguish student-users who revisit the same collection of Websites session-to-session from student-users who extend their Website *repertoire* by visiting previously unvisited Websites. The repertoire is the cardinality of the *vocabulary* set where the vocabulary set is the set of different Websites (or different Webhosts). Repertoire and vocabulary are discussed in Chapter three.

It might be thought that student-users who have a large average session click rate also have a large Website re-request rate. That is, at least in part, a large Website re-request rate generates a large session click rate (despite, in general, the increase in click rate not being explained by the increase in Website re-requesting). Figure 4.6 illustrates the relationship between these two user-characterizations during study-year two[4] which reveals that the association between them appears less strong for large average session click rates.

[4] The corresponding scattergram in respect of study-year one is illustrated in Figure C.2 in Appendix C.

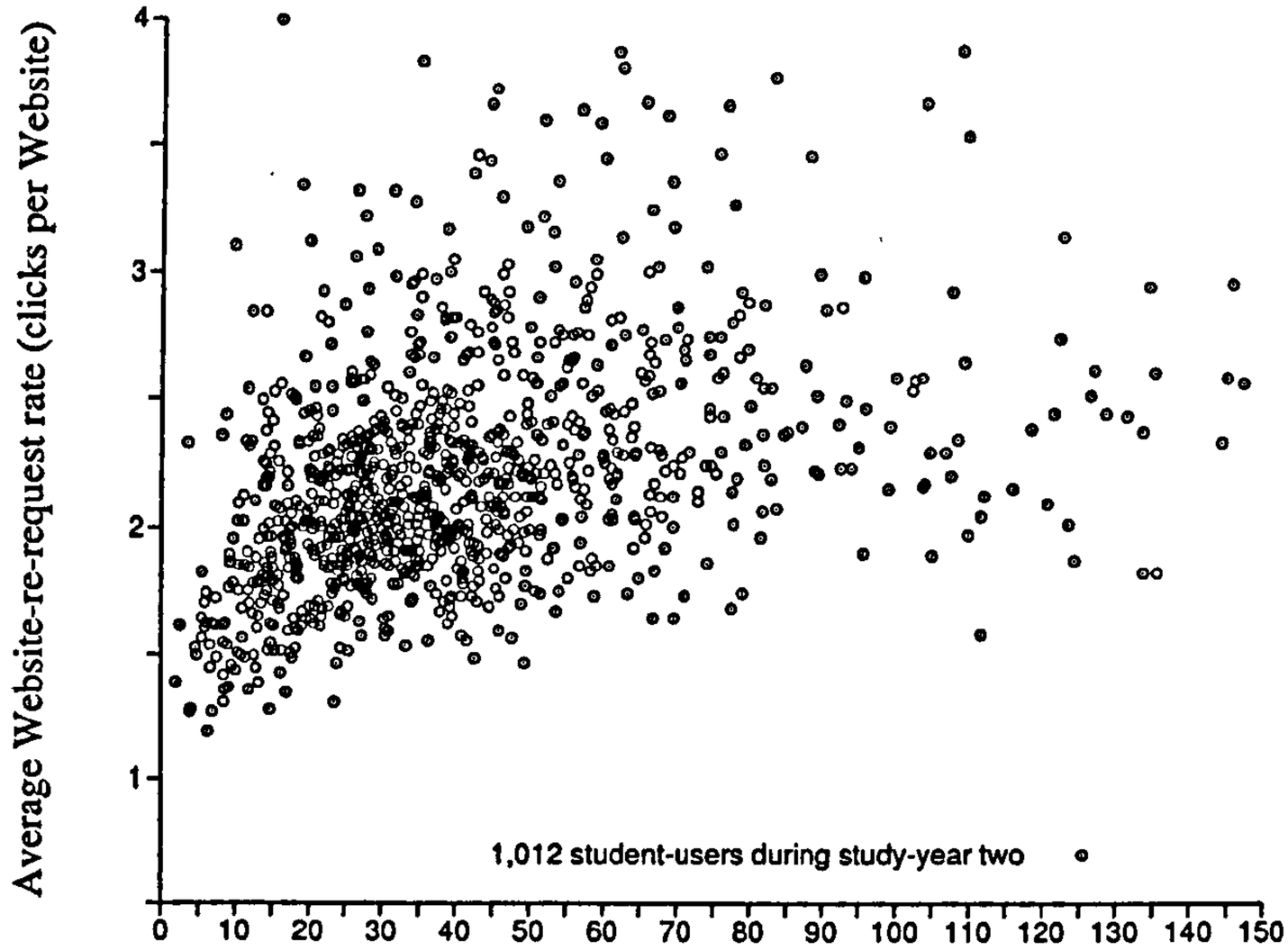

Figure 4.6: Scattergram of 1,012 student-user's average session click rate and average Website-re-request rate during study-year two (range illustrated up to 150 clicks per session and four clicks per Website)

1,012 out of the 1,050 student-users are represented in this scattergram; 269 have an average session click rate greater than 50 clicks per session that is they appear in the right-hand portion while the remaining 743 student-users appear in the left-hand portion. The linear correlation coefficients for these two portions are (left) Pearson $r = 0.350$ and (right) Pearson $r = 0.028$ which confirms the visual appearance that the association between average session click rate and average Website re-requesting rate is weaker towards the right.

During study-year one there is a similar apparent difference in the association between average session click rate and average Website re-requesting rate user-characterizations for left and right-hand portions of the scattergram partitioned as just described and the 927 student-users towards the left have $r = 0.519$ while the 115 on the right have $r = -0.135$.

This reinforces the suggestion that student-users who have large average session click rates differ from their colleagues in how they locate Web information. Moreover, in respect of how student-users with small average session click rates locate Web information, there is a possibility that average Website-re-request rate may be associated with average session click rate (since during both study-years one and two the correlation (Pearson r) is significant, ($p < 0.05$).

Average Webhost-persistence

During study-year one, student-user's average session Webhost-repertoire (that is the average for each student-user of the number of different Webhosts visited during each session) is 6.1(0.13) Webhosts per session. This increases ($p < .001$, $z = 10.2$) to 8.2(0.16) Webhosts per session during study-year two.

Figures C.4 and C.5 in Appendix C illustrate the average session Website and Webhost repertoire frequency distributions. In isolation neither the average session Website-repertoire nor the average session Webhost-repertoire is a reliable characterization because the duration of the session is not known.

The ratio of the number of different Websites which are visited to the number of different Webhosts gives the *Webhost-persistence* metric which is measured in Websites per Webhost. This metric is similar in principle to the *path length* metric which is discussed in Chapters two and three. The Webhost-persistence of a student-user during a session is,

$$\text{session Webhost-persistence} = \frac{\text{session Website-repertoire}}{\text{session Webhost-repertoire}}, \text{ (Websites per Webhost).}$$

The *average*[5] *Webhost-persistence* user-characterization of a student-user is,

$$\text{average Webhost-persistence} = \frac{\sum \text{session Website-repertoire}}{\sum \text{session Webhost-repertoire}}, \text{ (Websites per Webhost).}$$

where the summation is over all the student-user's sessions during the study-year.

Hence, session-by-session, the Webhost-persistence indicates the extent to which a student-user is locating his (or her) Web information seeking in several Websites within a Webhost or visiting just a single Website within a Webhost. As with the Website-re-requesting and the Website-trajectory slope user-characterizations, the Webhost-trajectory slope characterization is needed to distinguish between those student-users who each session visit Webhosts which they have previously visited at some time from student-users who increase their Webhost repertoire.

During study-year one the mean student-users' average Webhost-persistence is 2.3(0.03) Websites per Webhost which increases ($p < .005$, $z = 2.8$) slightly to 2.4(0.02) Websites per Webhost during study-year two. Figure 4.7 illustrates the frequency distributions which shows the heavy tail phenomenon. For example, during study-year two, 19 student-users or 1.8% of the distribution exceed the $99\frac{9}{10}^{\text{th}}$ centile

[5] As before, this average is not an arithmetic mean but it can be interpreted as an average gradient.

expected[6] of a Gaussian distribution in that their average Webhost-persistence > 4.49 (= 2.36 + 3.09 × 0.69) Websites per Webhost.

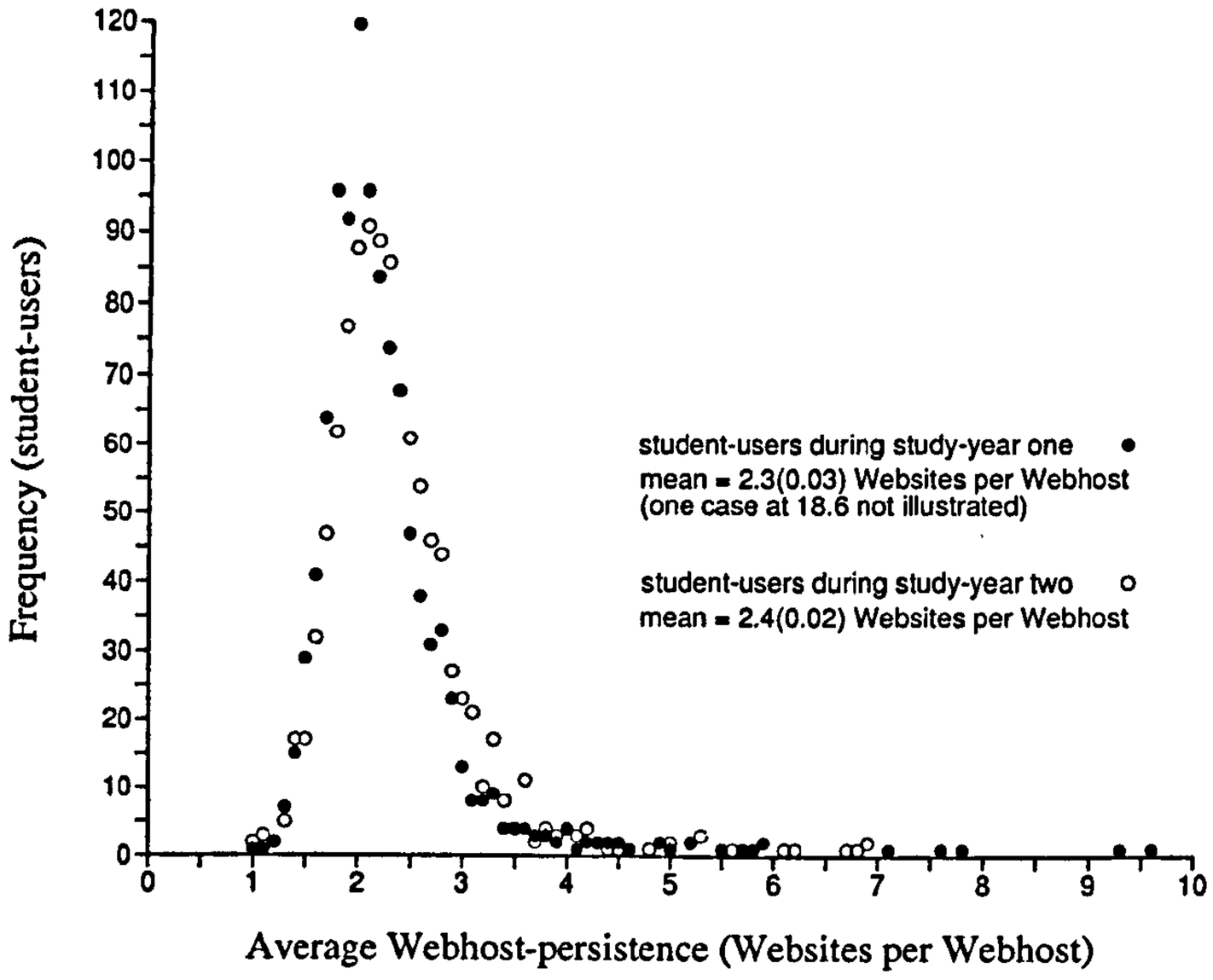

Figure 4.7: Frequency distributions of student-user's average Webhost-persistence

The apparent differences in the association between the average session click rate and the average Webhost-persistence user-characterizations for student-users repeats the apparent differences in the association between the average session click rate and average Website-re-request rate user-characterizations. That is, the association appears less strong for student-users who have large average session click rates. This is shown in respect of study-year two[7] by the scattergram of average session click rate and average Webhost-persistence which is illustrated in Figure 4.8.

[6] The calculation has not been modified to take account of the truncation at zero but in this case the effect is not material.

[7] The equivalent scattergram in respect of study-year one is illustrated in Figure C.6 in Appendix C.

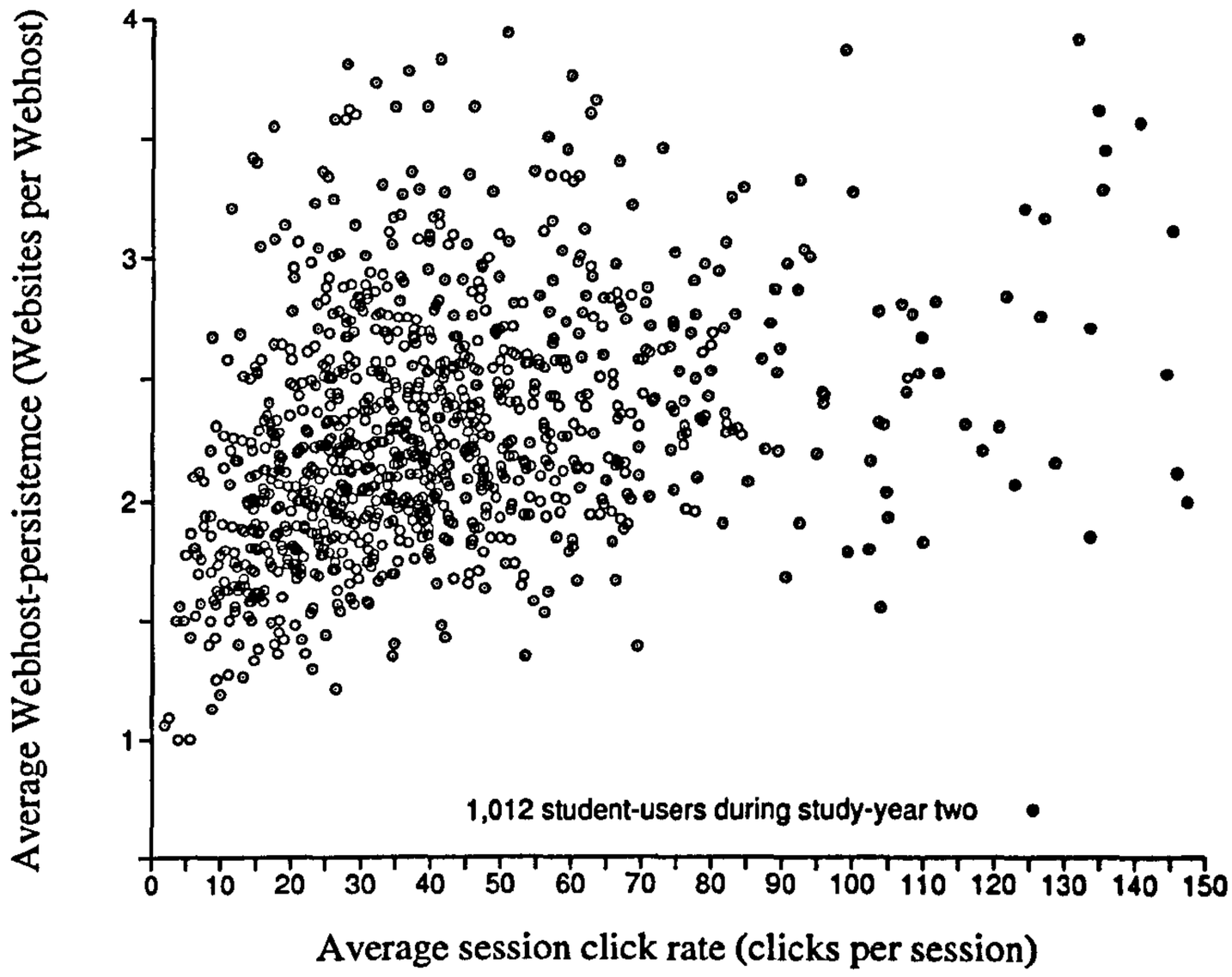

Figure 4.8: Scattergram of 1,012 student-user's average session click rate and average Webhost-persistence during study-year two (range illustrated up to 150 clicks per session and four Websites per Webhost)

As before, the visual impression is supported by the linear correlation coefficients of the left ($\leqslant$ 50 clicks per session) and right-hand portions of the scatter which are $r = 0.359$ for 734 student-users and $r = 0.171$ for 278 student-users respectively. During study-year one the equivalent linear correlation coefficients are $r = 0.319$ for 919 student-users and $r = -0.023$ for 102 student-users respectively.

Since both the average Website-re-request rate user-characterization and the Webhost-persistence user-characterization relate similarly to the average session click rate user-characterization then it might be conjectured that there is a strong relationship between a student-user's Webhost-persistence and Website-re-request rate user-characterizations. The scattergram in respect of 1,006 student-users during study-year two[8] is illustrated in Figure 4.9. However there is no linear correlation generally between student-user's average Website-re-request rate and average Webhost-persistence user-characterizations (Pearson $r = 0.049, p > .05, n = 1,050$). During study-year one there is a small linear correlation (Pearson $r = 0.062, p < .044, n = 1,050$) which, although a significant correlation at $\alpha = .05$ explains less than 4% of variability. The conjecture therefore fails.

[8] The equivalent scattergram in respect of study-year one is illustrated in Figure C.8 in Appendix C.

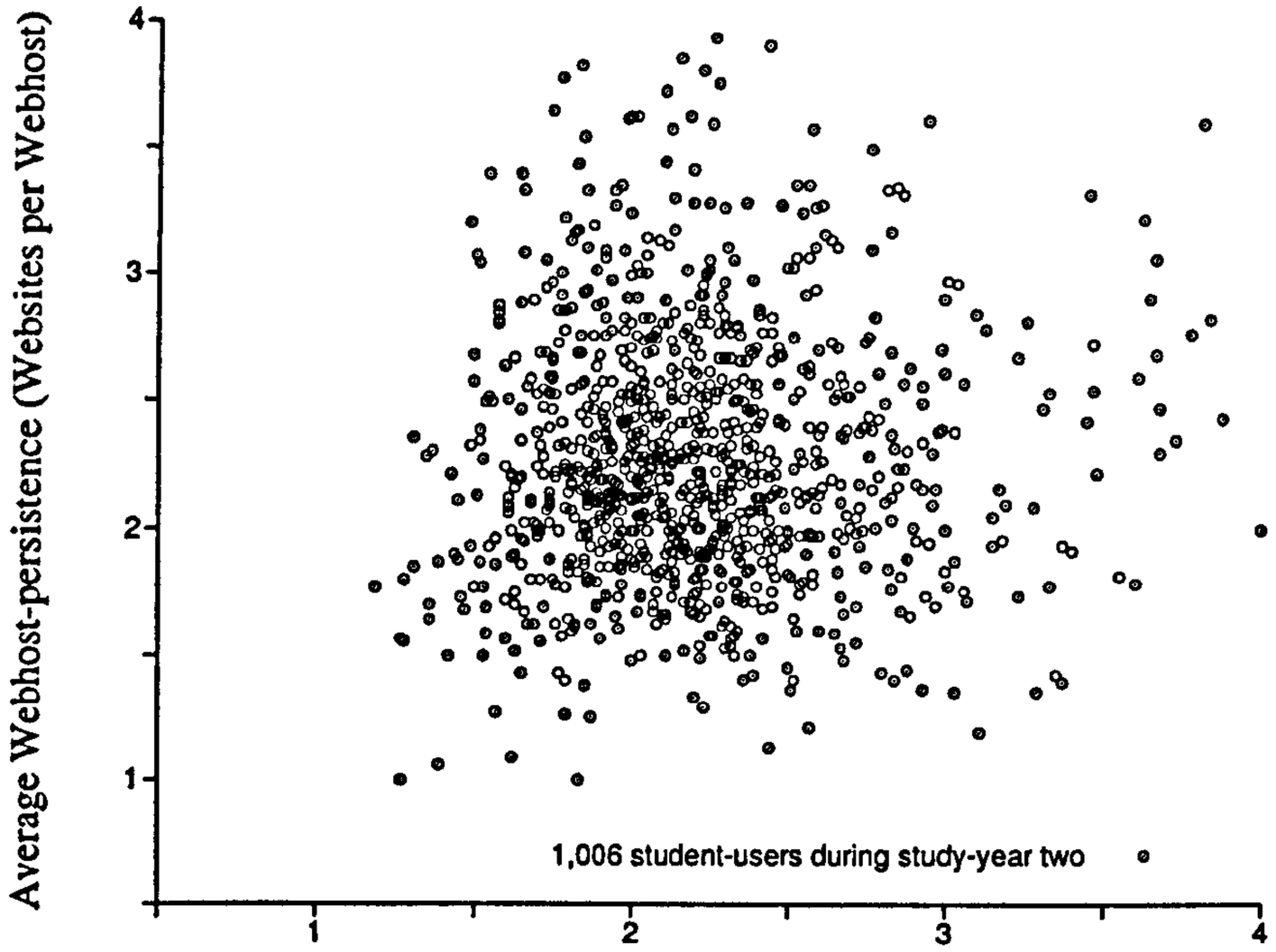

Figure 4.9: Scattergram of 1,006 student-user's average Website-re-request rate and average Webhost-persistence during study-year two (range illustrated up to four clicks per Website and four Websites per Webhost)

Thus, overall, student-users who *revisit* the same Website during a session are no more or less likely to visit other Websites at the Webhost (and thereby increase their Webhost-persistence) than are other student-users.

All of the user-characterizations which have been used so far are session-by-session user-characterizations. They therefore describe and differentiate each user only in respect of their information seeking activity within a single session.

It has already been noted that when the two study-years are compared student-users appear to become more energetic, that is, they click more during each session (but this may be due to changes in the duration of sessions). They also become more active, that is, the proportion of query-clicks is greater. The Website-re-request rate increases but not sufficiently to account for all the increase in session Website-repertoire. Thus during each session student-users visit more different Websites and revisit more Websites. The session Webhost-repertoire rate also increases as does Webhost-persistence so that student-users visit more different Webhosts and more different Websites within each Webhost during each session.

These characteristics can be related by an approximate session-arithmetic as, for example,

average session click rate =

average session Website-re-request rate × average Webhost-persistence

× average session Webhost-repertoire.

Thus, during study-year one,

29.0 (clicks per session) ≈ 28.06

= 2.0 (clicks per Website) × 2.3 (Websites per Webhost) × 6.1 (Webhosts per session)

while, during study-year two,

44.1 (clicks per session) ≈ 45.264

= 2.3 (clicks per Website) × 2.4 (Websites per Webhost) × 8.2 (Webhosts per session).

Hence the average *magnitude* of each session is greater or comprises more Websites and more Webhosts during study-year two compared to during study-year one.

The session-arithmetic only considers the average arithmetic relationship of one user-characterization to another. It does not address how user-characterizations vary either amongst or between themselves. Nor does it address how student-users locate Web information session-to-session. Session-to-session user-characterization are discussed below. The average session-conformance user-characterization, also discussed below, is a session-by-session user-characterization but it indicates some of the variation in how student-users locate Web information. This shows that even though the magnitude of sessions is greater, each student-user's sessions become more not less alike one to another in respect of the Websites which the student-users visits during each session.

The session-arithmetic could be normalised by dividing each metric by the session click rate. This would then describe users' Web information seeking *by-click* rather than *by-session* and thus eliminate any dependency on session duration. In consequence each user could be described by an information seeking *profile* which would allow investigation of how scalar is student-users Web information seeking. That is, are the session Website-repertoire and session Webhost-repertoire of a session with

many clicks a scalar multiple of the session Website-repertoire and session Webhost-repertoire of a session with few clicks? If this is not the case then how do repertoires vary with clicks? The distribution in respect of study-year two[9] illustrated in Figure 4.10 suggests that student-users may be categorised according to how their relative preference for *skimming* over the Web visiting many Webhosts but not delving into those Webhosts, or *diving* into the Web by visiting fewer Webhosts but many Websites within those Webhosts. Profile analysis falls outside the scope of this investigation but is discussed briefly in Chapter seven.

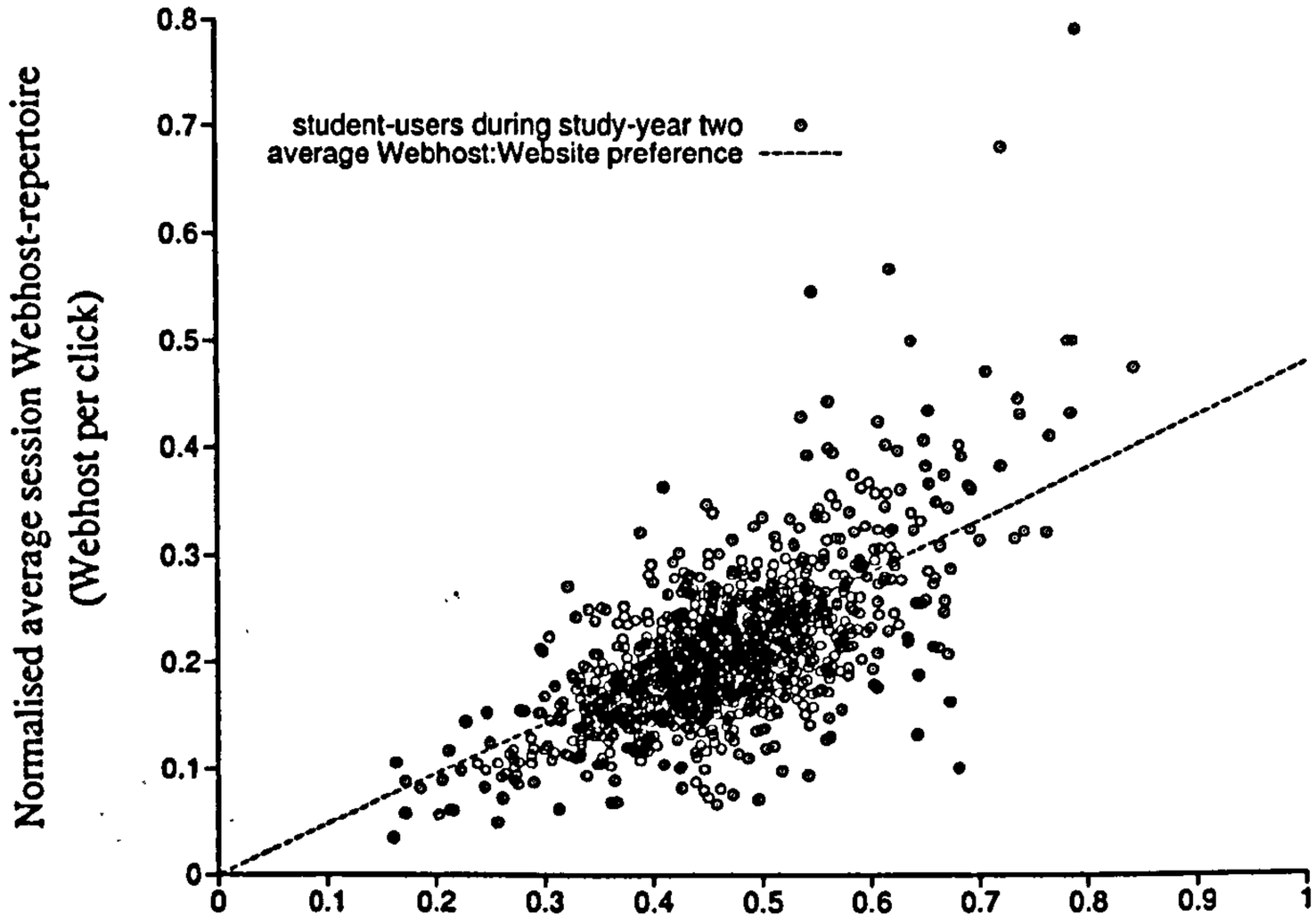

Figure 4.10: Scattergram of student-user's click normalised average session Website-repertoire and average session Webhost-repertoires during study-year two

Website-trajectory slope

Trajectory functions are discussed in Chapter three. The *Website-trajectory* for a student-user, τ_{Website}, relates the student-user's Website-repertoire and click rate during a study-year, that is,

$$\text{Website-repertoire} = \tau_{\text{Website}}(\text{cumulative click frequency})$$

so that the *Website-trajectory slope*, measured in Websites per click, characterizes the growth of the student-user's Website-repertoire during the study-year.

[9] The equivalent scattergram in respect of study-year one which has the same appearance is illustrated in Figure C.10 in Appendix C.

In a similar way the *Webhost-trajectory slope*, measured in Webhosts per click, characterizes the growth of a student-user's Webhost-repertoire. The Webhost-trajectory is constrained above by the Website-trajectory so the Webhost-trajectory slope $\leqslant$ Website-trajectory slope.

The trajectory function, τ is modelled linearly so that $\tau(n) = Tn + constant$ where $0 \leqslant T \leqslant 1$ and n is the cumulative click frequency.

Both the Website-trajectory slope and Webhost-trajectory slope characterizations are session-to-session characterizations. Thus, for example two student-users who session-by-session each submit twenty Web requests to five different Websites have the same average session Website-repertoire (= 5 Websites per session) and same average Website-re-request rate (= 4 clicks per Website). But if the first student-user during each subsequent session revisits only two Websites visited during a previous session compared with four by the second student-user then the first will have a Website-repertoire growth rate of $\frac{3}{20}$ Websites per click compared to only $\frac{1}{20}$ Websites per click for the second student-user. How each locates Web information session-to-session is thereby differentiated.[10]

Figure 4.11 illustrates Website-trajectory function graphs constructed for each student-user. The graphs are truncated artificially at 5,000 clicks in order to retain detail within the illustration. The Website-trajectory slope characterizes each student-user by the gradient of the straight line fitted to each graph. A student-user with a Website-trajectory slope user-characterization of 0.3 Websites per click therefore has a Website-trajectory function which is approximated by the straight line,

$$\text{Website-repertoire} = 0.3 \times \text{cumulative Web requests.}$$

Thus after 1,000 clicks this student-user is expected to have visited 300 different Websites.

[10] The trajectory slope user-characterizations here would not be exactly 0.15 and 0.05 Websites per click because of the initial conditions and the straight line fitting procedure used.

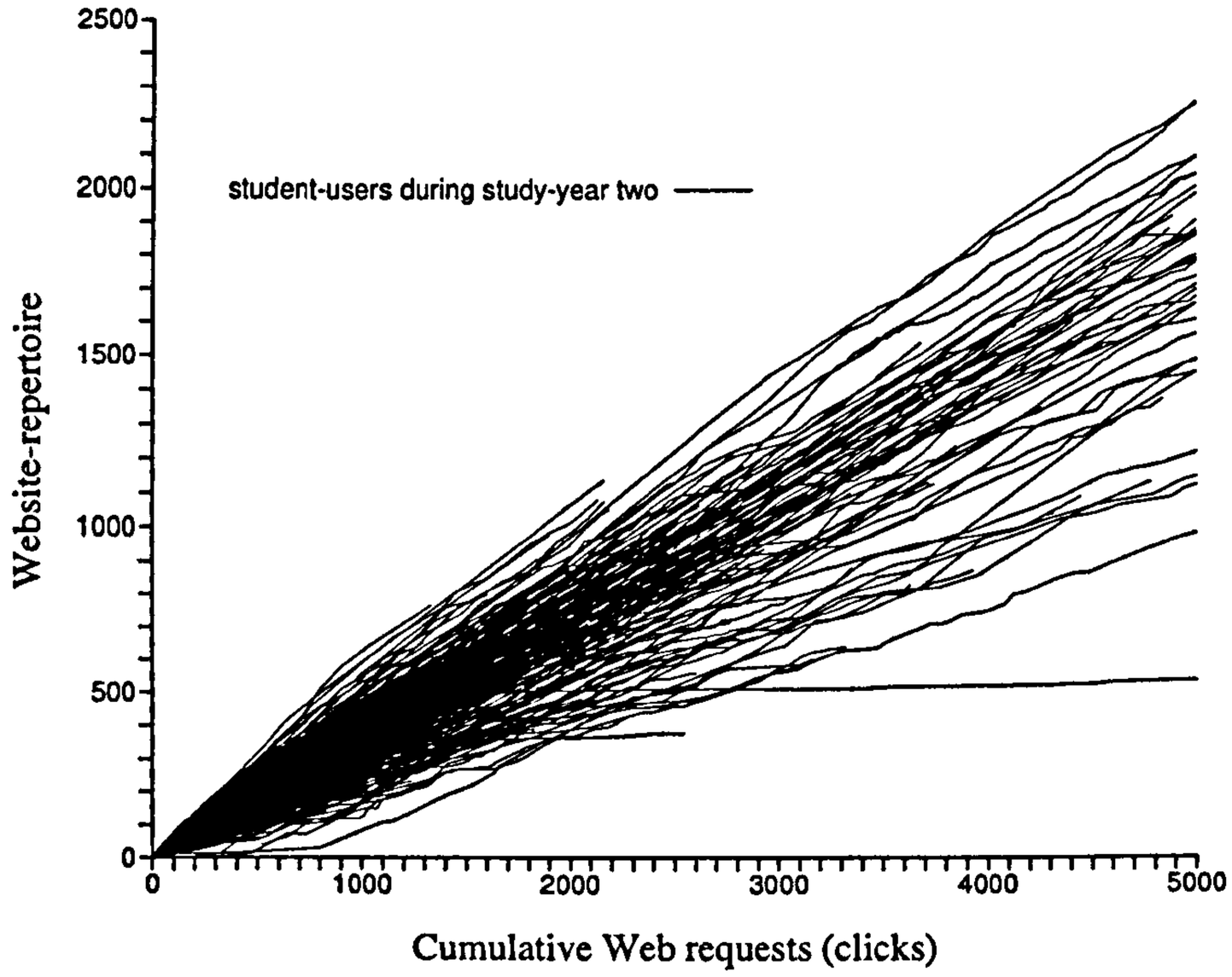

Figure 4.11: Website-trajectory function graphs for student-users during study-year two (range illustrated up to 5,000 clicks)

Since for each student-user the Webhost-repertoire is constrained above by the Website-repertoire, growth in the Webhost repertoire is slower but the overall features of the Webhost-trajectories are the same as the Website-trajectories. Trajectory graphs are illustrated in Figures B.1 to B.4 in Appendix B.

The mean Website-trajectory slope user-characterization for student-users during study-year one is 0.40(0.004) Websites per click. This reduces ($p < .001$, $z = 4.0$) to 0.38(0.003) Websites per click during study-year two. This session-to-session reduction is compatible in principle with the greater magnitude seen above in the session-by-session user-characterizations and the hypothesis that student-users' session-to-session and session-by-session Web information seeking is similar. That is, generally, when more different Websites are visited within a session then a similar proportion of more different Websites are visited overall. Figure 4.12 illustrates the frequency distribution of the Website-trajectory slope user-characterizations during each study-year.

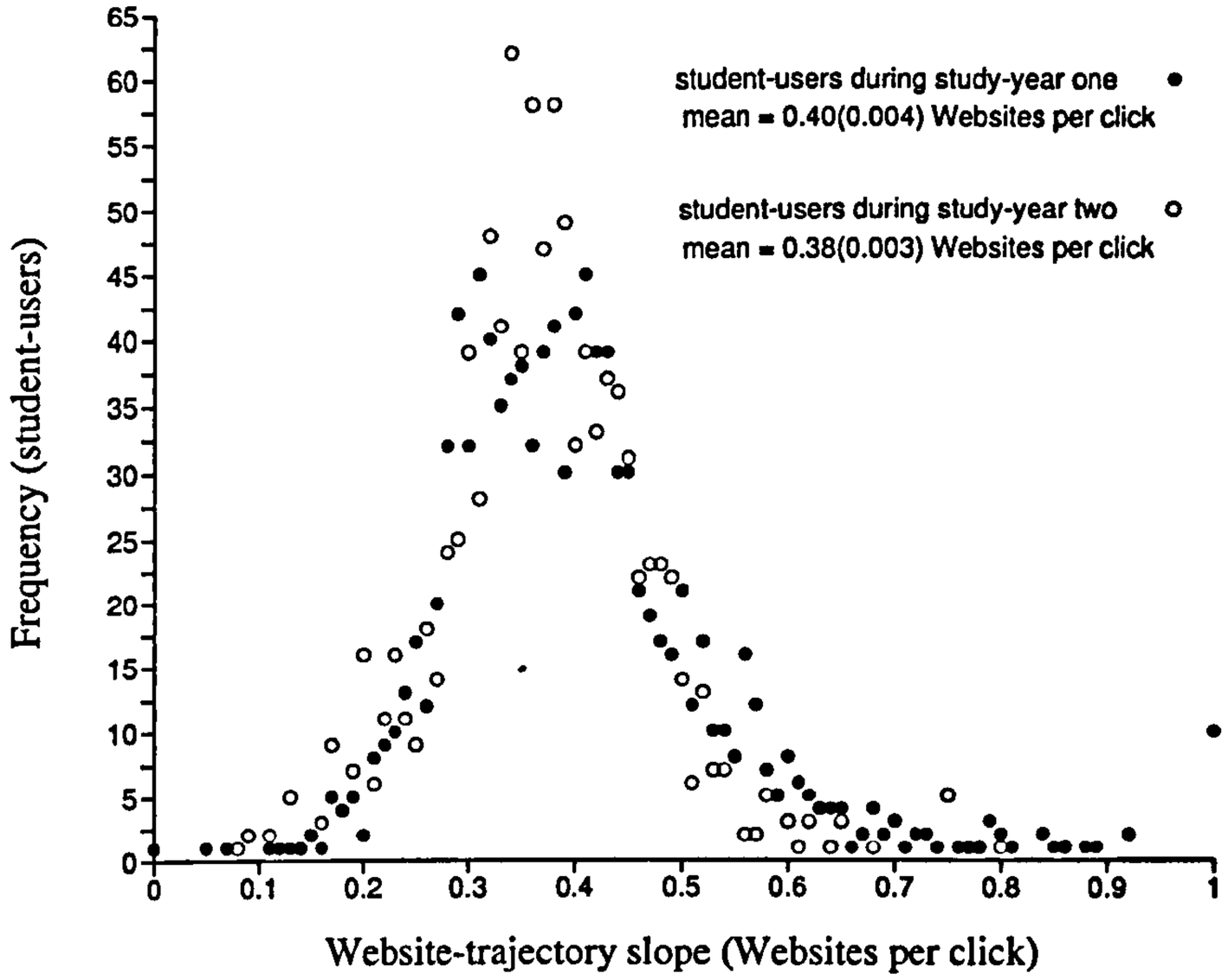

Figure 4.12: Frequency distributions of student-user's Website-trajectory slope

Figure 4.13 illustrates the frequency distributions of the Webhost-trajectory slopes corresponding to the Website-trajectory slope user-characterizations. The reduction ($p < .001,\ z = 5.5$) in the mean Webhost-trajectory slope for student-users is from 0.16(0.003) Webhosts per click during study-year one to 0.14(0.002) Webhosts per click during study-year two. Hence by a similar argument as previously when more different Webhosts are visited within a session then equivalently more different Webhosts are visited overall (and vice versa).

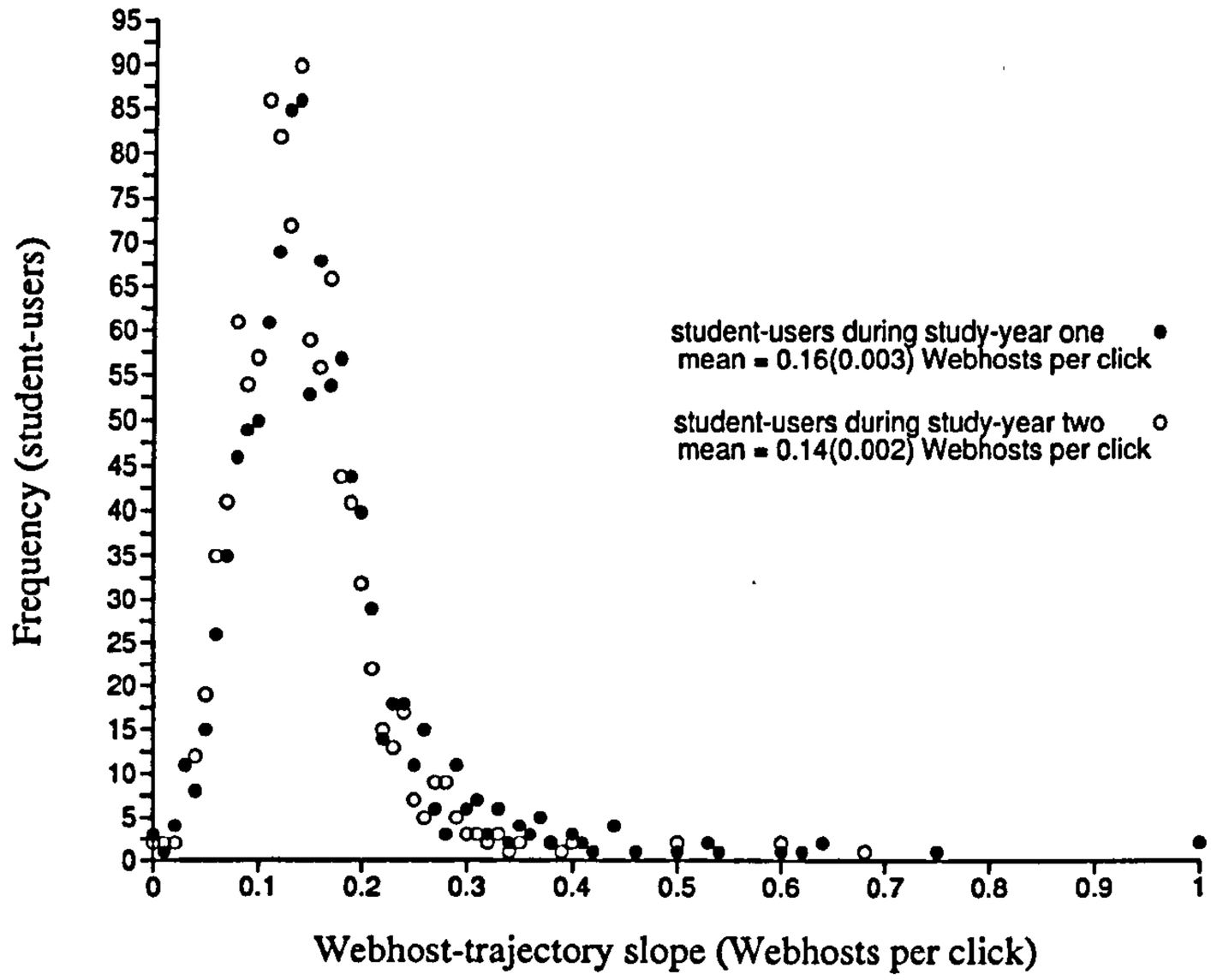

Figure 4.13: Frequency distributions of student-user's Webhost-trajectory slope

The general similarity in how student-users locate Web information session-by-session and session-to-session is demonstrated by comparing an actual session-by-session metric with the predicted value derived from the session-to-session metrics. During study-year one, session-to-session, student-users grow their Website and Webhost repertoires by 0.40 Websites per click and 0.16 Webhosts per click respectively. Their predicted longer term Webhost-persistence during study-year one is thus $\frac{0.40}{0.16} = 2.5$ Websites per Webhost. The predicted value for study-year two is an increase to 2.7 ($= \frac{0.38}{0.14}$) Websites per Webhost.

The session-by-session average session Webhost-persistence actually increased from 2.3 to 2.4 Websites per Webhost. The good general agreement here suggests that the session-by-session user-characterizations provide a reliable indication of session-to-session Web information seeking. That is, how students locate Web information session-by-session can be generalised to how students locate Web information session-to-session. Thus, the profile of the combination of two sessions from a student-user is similar to the separate profiles of each of the sessions. (Profile analysis is noted as an area of possible future work. In addition to differences in profile between student-users there are indications that some student-users change their profiles while others retain their profiles.)

The alternative, that student-users visit and revisit Websites and Webhosts during each session in a particular proportion but then, in subsequent sessions visit an entirely different collection of Websites is dismissed.

Average session-conformance and session-conformance range

Both these user-characterizations are based on the *session-conformance* metric which is computed for each session. The average session-conformance user-characterization is computed for each student-user as,

$$\text{average session-conformance} = \frac{\sum \text{session-conformance}}{\text{session rate}}$$

$$= \frac{1}{n} \sum_{\text{sessions}=1}^{n} \text{session-conformance}$$

and the average session-conformance user-characterization for a student-user for each study-year is found by computing the summation of the sessions over one study-year only.

Each student-user's session-conformance range user-characterization is,

$$\text{session-conformance range} = \text{session-conformance}_{\,maximum} - \text{session-conformance}_{\,minimum}$$

where the session-conformances refer to the student-user's sessions during each of the study-years.

The session-conformance metric measures how much one Web information seeking session resembles (or more accurately is dissimilar to) another. This is in a lexical rather than a semantic sense.[11] At the extreme, if two sessions comprise Web requests to the same Websites in the same proportions then the two sessions are regarded as being the same. In this case the session-conformance metric would be zero, that is the *smaller* the metric the *greater* the resemblance.

Computation of the session-conformance metric, which uses of the vector model of information retrieval, is described in Chapter three. The notion of *conformance* comes from this metric measuring the extent to which a particular session, or group of sessions, conforms to the average session in respect of the collection of visits to Websites. Analysis of the Web log generates the Website vocabulary during both study-years. Hence, in principle, each of the 46,558 sessions can be represented as a position vector or point in a Website space where the co-ordinates of the point are the frequency of occurrence of the Websites in the session. The position vector of each session can be normalised to be of unit length so the normalised points are distributed over a (many dimensional) unit quadrant. The normalised centroid of all the session points is also a point on the unit quadrant and represents the average

[11] That is, a session is a bag of Websites; the substance of Websites is not considered.

session. Hence the displacement of each session from the centroid can be computed. The detail of the computation ensures that every different session (that is sessions which are not the same) has a different displacement and that the displacements of sessions which resemble each other are numerically close.

In practice, for numerical convenience, the session-conformance metric is the adjusted squared displacement of the normalised session point from the normalised centroid. The displacement is the usual Euclidean distance and the adjustment reduces each displacement by a fixed amount so that the session-conformance metric runs from zero, for those sessions which most resemble the average session, up to about 1.6 for those sessions which least resemble the average session.

The weighting in the calculation of the displacement in the Website space for the session-conformance metric gives more attention to those Websites which occur more frequently in sessions. This is achieved in part[12] by ranking each Website according to its session frequency and considering a Website space of just one thousand dimensions corresponding to the one thousand most frequently occurring Websites plus one extra dimension for all the other Websites, rather than a Website space of the complete 507,618 Website repertoire.

Taken together therefore the average session-conformance user-characterization and session-conformance range user-characterization describe and differentiate how student-users locate Web information in respect of how much the collection of Websites which they visit and revisit during each of their sessions resembles the overall average session. If a student-user's session Website vocabulary is always the same and these Websites are always visited in the same proportions his (or her) session-conformance metric will be constant. Hence this student-user's average session-conformance user-characterization is whatever value this constant is and his (or her) session-conformance range user-characterization is zero. The smaller the average session-conformance user-characterization the greater the resemblance to the overall (both study-years) average session. Two student-users could each have the same average session-conformance user-characterization but one may never vary[13] his (or her) sessions while the other has sessions with session-conformance greater and smaller than his (or her) average session-conformance user-characterization.

That is, student-users with a small average session-conformance user-characterization resemble each other more in how they locate Web information than do student-users with larger average session-conformance user-characterizations, and how a student-user with a larger session-conformance range user-characterization locates Web information session-to-session is more eclectic (that is the Websites visited are from

[12] A tf*idf weighting scheme is also used, see Chapter three.
[13] In the sense of the normalised ranked Website space of 1,000 different Websites plus one.

a larger vocabulary) than a student-user with a smaller session-conformance range user-characterization.

The mean average session-conformance user-characterizations during study-years one and two are 1.15(0.006) and 1.11(0.007) respectively. The reduction is significant ($p < .001,\ z = 4.3$). Hence, overall, student-users resemble one another more in how they locate Web information during study-year two than during study-year one (since the mean average session-conformance user-characterizations indicates sessions closer to the overall average session). The frequency distribution of the average session-conformance user-characterization is illustrated in Figure 4.14.

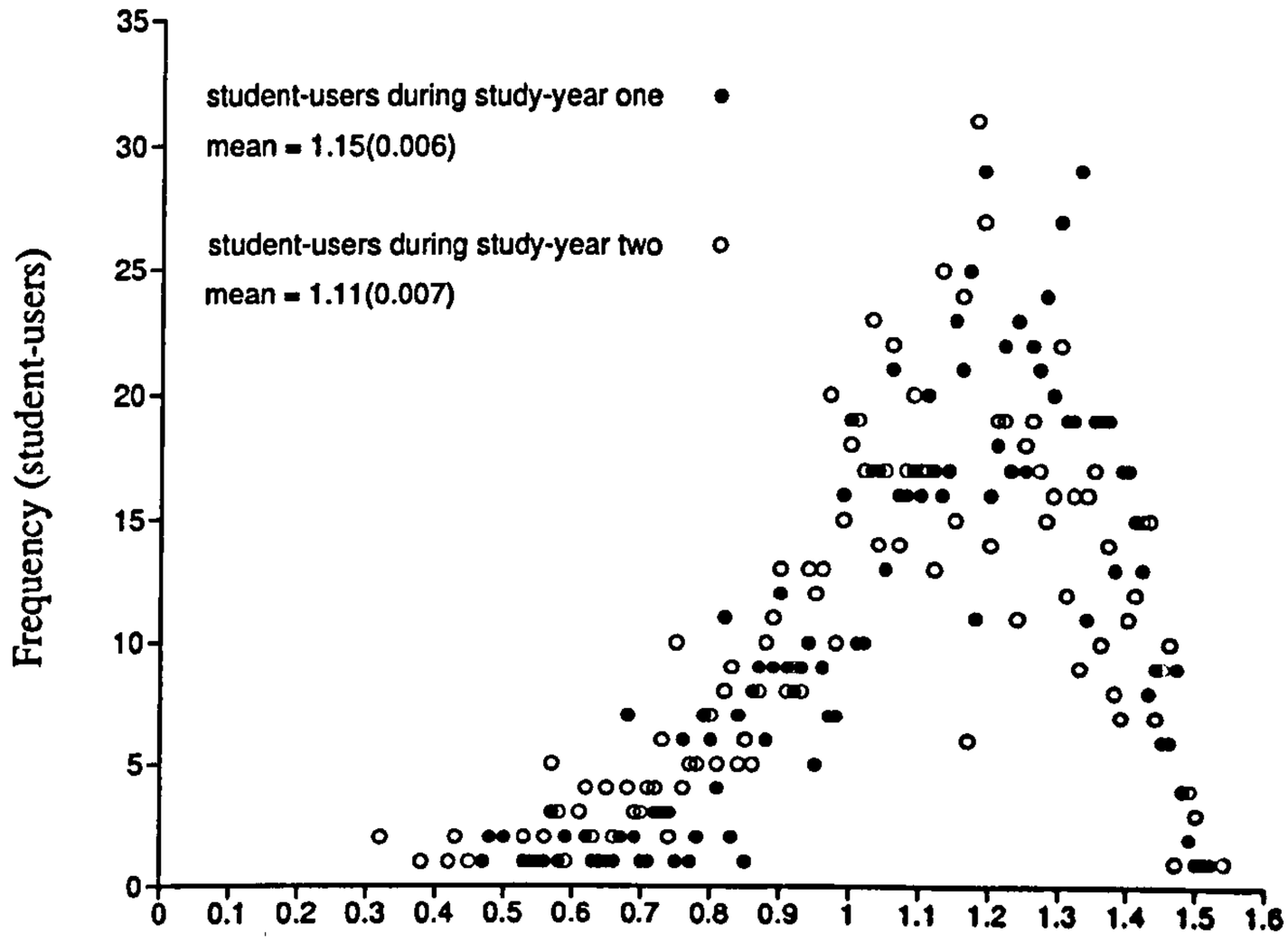

Figure 4.14: Frequency distributions of student-user's average session-conformance

The frequency distribution of the session-conformance range user-characterization is illustrated in Figure 4.15. The mean value of the user-characterization overall increased ($p < .001,\ z = 5.4$) from 1.06(0.018) during study-year one to 1.19(0.016) during study-year two. This indicates that overall student-users are more eclectic as regards their session Website-vocabulary during study-year two compared to during study-year one. The frequency distribution graph shows that the session-conformance range user-characterization distribution consists of two parts. This is explained as follows. A session which comprises exclusively visits to rare Websites (2,674 sessions during study-year one and 3,605 sessions during study-year two) receives a session-conformance metric of zero. Therefore student-users who sometime during a study-year undertake such a session have a minimum session-conformance of

zero. Those who don't have a minimum session-conformance > 0.478. The maximum session-conformance is about 1.554. Therefore student-users who do not have sessions which consist exclusively of visits to rare Websites can have a range of session-conformance which is no more than about $(1.554 - 0.478 \approx 1)$. Hence student-users whose Web information seeking never includes a session which exclusively comprises rare Websites appear on the left of the graph. Analysis shows that the student-users who appear in the left of the graph are just those student-users whose Web information seeking never includes a session which exclusively comprises rare Websites since it is found that student-users whose minimum session-conformance is zero also have a session-conformance range user-characterization > 1.

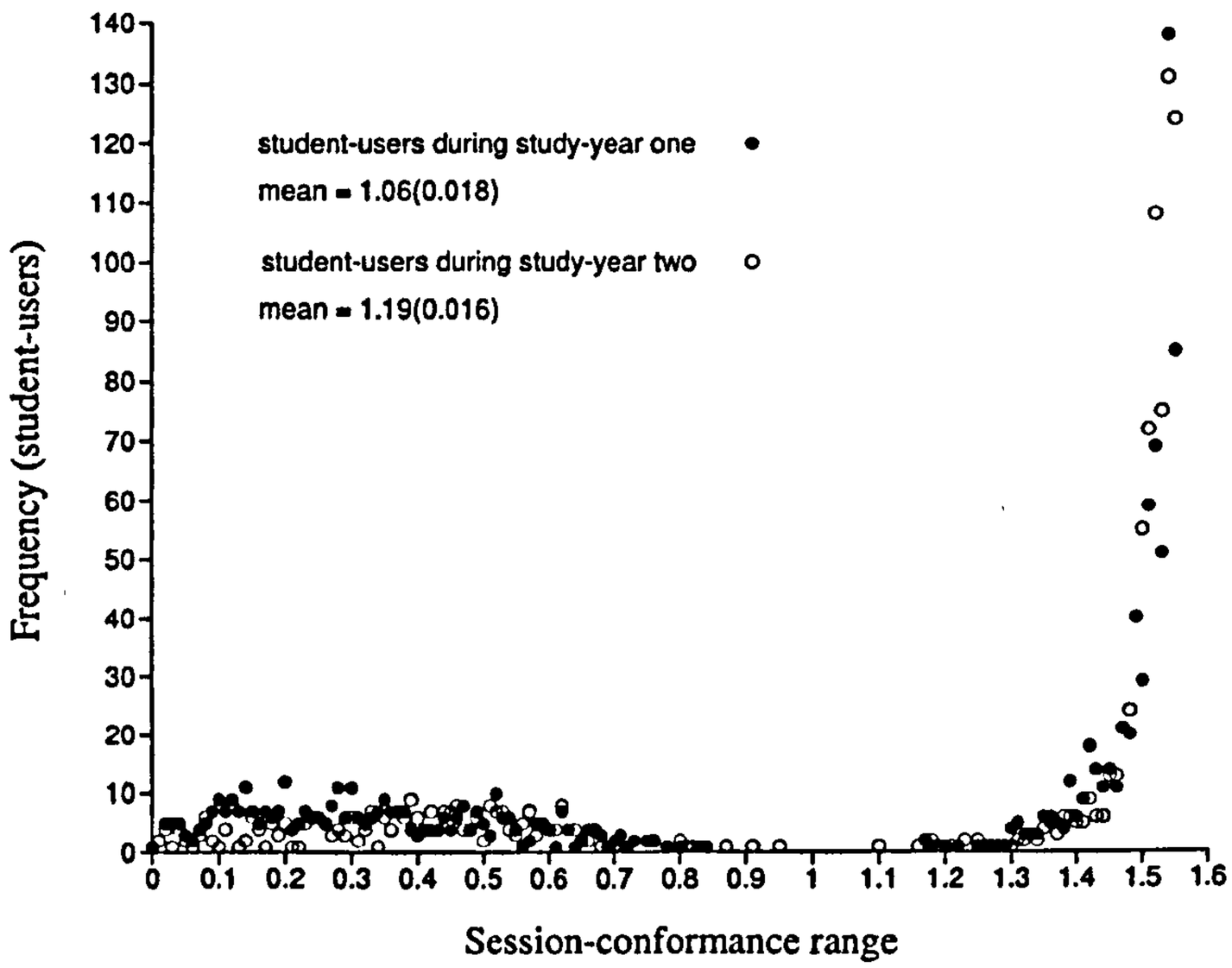

Figure 4.15: Frequency distributions of student-user's session-conformance range

Figure 4.16 shows, for study-year two,[14] the relationship between a student-user's average session-conformance user-characterization and session-conformance range user-characterization. The two clusters of student-users illustrate the argument explaining the bipartite distribution shown in Figure 4.15. The lower cluster is student-users with a small session-conformance range, that is precisely those students who never have a session consisting exclusively of visits to rare Websites. In consequence their minimum session-conformance > 0.478 and hence their average session-conformance $\not< 0.478$ which pushes them over to the right of the graph. (The cluster is much further to the right than this which indicates a greater proportion of sessions that include high ranking Websites.)

[14] The equivalent graph in respect of study-year one is illustrated in Figure C.12 in Appendix C.

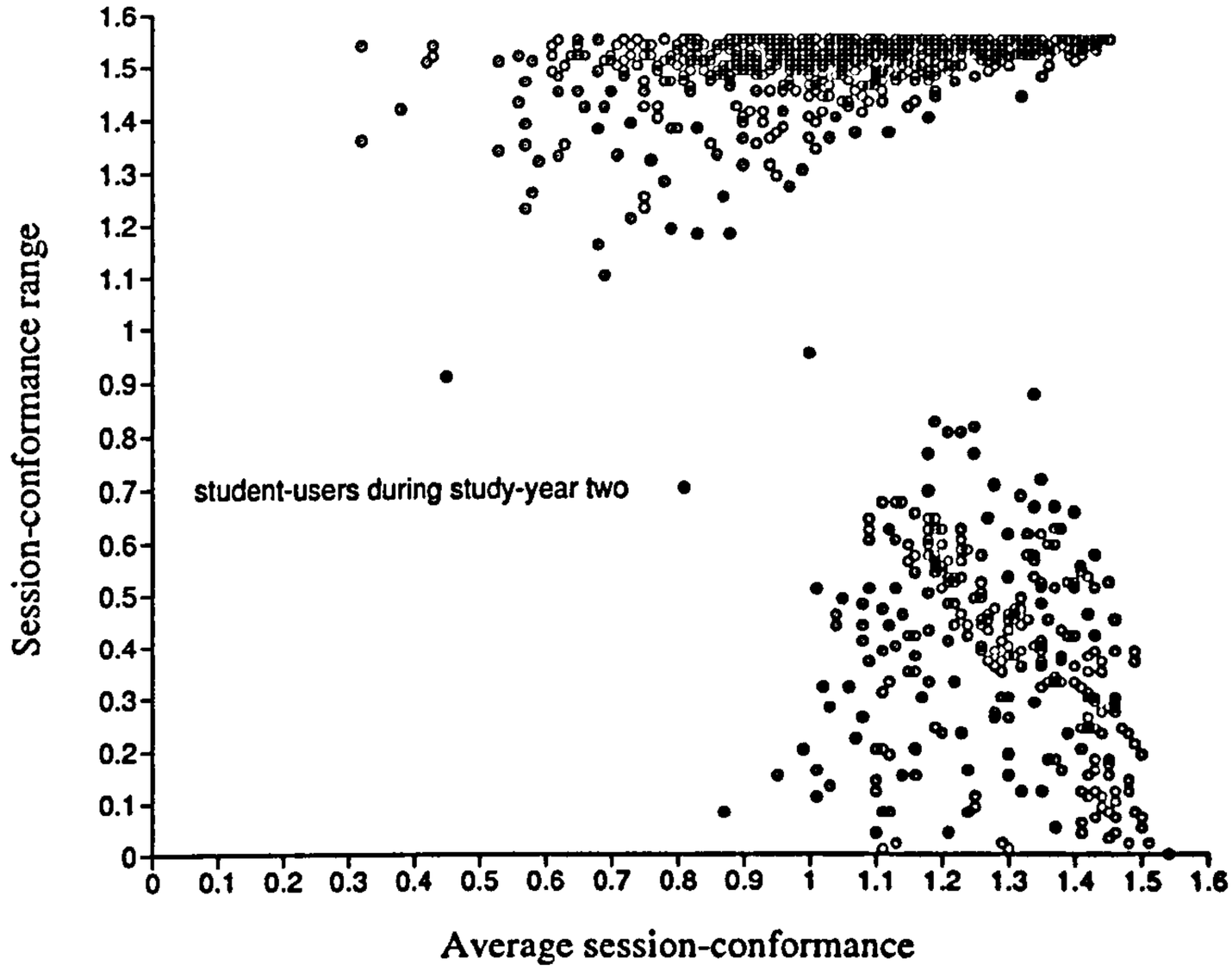

Figure 4.16: Scattergram of student-user's average session-conformance and session-conformance range during study-year two

The upper cluster is all those student-users who do have a session consisting exclusively of visits to rare Websites. The larger spread in average session-conformance combined with a smaller spread of session-conformance range for these student-users indicates that they do not much resemble each other in how they locate Web information, but how each student-user locates Web information during a session more strongly resembles another of that student-user's sessions.

Since whether or not a student-user has a session which is exclusively visits to rare Websites generates a dichotomous partition of student-users then this defines a user-attribute. Student-users are categorised as either *eclectic* if they have such a session or *conformant* otherwise. Eclectic student-users are doubly 'eclectic' in how they locate Web information. Firstly, by definition, they have sessions which are visits exclusively to rare Website and since the session frequency of each of these Websites is less than 41 but there are 506,618 of them, each student-user is sparingly visiting a few Websites from a large selection. Secondly, how eclectic student-users locate Web information is collectively more varied but individually more distinctive since they have a broader spread of average session-conformance user-characterizations than conformant student-users. Hence eclectic student-users are 'eclectic' at a session level as well as at the level of the Websites in each session.

On the other hand conformant student-users always visit (at least one of) those Websites which are collectively the top one thousand most visited (by-session) Websites.

Their selection of Websites thereby conforms to the norm. Associated with this, the scope for individual variation in the Websites visited is reduced hence the spread of average session-conformance for conformant student-users is smaller compared to that of eclectic student-users. How conformant student-users locate Web information resembles how other conformant student-users locate Web information more than the resemblance among each conformant student-user's sessions.

Table 4.1 reports the student-user frequencies for the eclectic and conformant user-attribute partitions during study-years one and two. The χ^2 test for consistency is $\chi^2 = 22$ which exceeds the critical value of $\chi^2_{1;0.005} = 6.63$. Hence the frequency of student-users who are eclectic/conformant is not consistent ($p<.005$) between each of the study-years. By inspection, the proportion of student-users who are eclectic is greater during study-year two. This is in line with an overall hypothesis that student-users become more eclectic over time but a study of individuals as in Chapter six is needed to consider the compensatory changes which may be occurring.

Study-year	**Conformance user-attribute**	
	eclectic	conformant
study-year one	655 student-users	395 student-users
study-year two	756 student-users	294 student-users

Table 4.1: Cross-tabulation of student-user frequency by study-year and conformance

In the next Section the analyses extend the initial description and differentiation of how student-users locate Web information by comparing groups of student-users defined by their user-attributes. Before that, Table 4.2 summarizes the results of this overall analysis based on the seven user-characterizations. For each user-characterization the differences between during study-year one and during study-year two are significant. Comparing during the two study-years, then how student-users locate Web information can be summarised as;

1. student-users become more energetic, that is they submit more Web requests (clicks),

2. they are more active, that is there is an increased proportion of query-clicking rather than link-clicking,

3. the magnitude, that is Website (and Webhost) session repertoire, of sessions increases,

4. there is also more Website revisiting and more visiting of different Websites within Webhosts,

5. on average sessions become more similar but student-users also become more eclectic so that individuals become more distinctive, and,

6. there is considerable variation so that most student-user's user-characterizations are overstated by the mean user-characterization because of the asymmetry of distribution which results from the underlying power law distribution of session-rate and session click rate metrics.

Table 4.2 summarizes the user-characterization metrics and the overall z test of difference statistic of the metrics between the study-years which are discussed in this Section. In the next Section the initial interpretation of the Web log which in this Section describes how student-users overall locate Web information is refined by considering separately groups of student-users defined by their student-user attribute.

Student-users by study-year	User-characterization metric						
	session-by-session					session-to-session	
	average session click rate	average query-click proportion	average Website-re-request rate	average Webhost-persistence	average session-conformance	Website-trajectory slope	session-conformance range
study-year one	29.0(0.85) clicks/session	0.26(0.003) qry-clicks/click	2.0(0.01) clicks/Website	2.3(0.03) Wsites/Whost	1.15(0.006)	0.40(0.004) Wsites/click	1.06(0.018)
study-year two	44.1(1.11) clicks/session	0.33(0.003) qry-clicks/click	2.3(0.02) clicks/Website	2.4(0.02) Wsites/Whost	1.11(0.007)	0.38(0.003) Wsites/click	1.19(0.016)
difference statistic	$z = 10.8$	$z = 14.0$	$z = 13.4$	$z = 10.2$	$z = 4.3$	$z = 4.0$	$z = 5.4$

Table 4.2: Cross-tabulation of overall user-characterizations by study-year

4.3 Similarities and differences

In this Section the overall description and differentiation of how student-users locate Web information discussed in the previous Section is extended and refined by comparing groups of student-users. These groups are defined by the user-attributes of (a) the student-user's gender (b) the student-user's session-rate, and (c) the student-user's conformance.

The anonymous codes representing student-users in the Web log and which allow the log to be analysed by-user can also be connected to identically coded student-user demographic information. In consequence the gender of each anonymous student-user is known (there are 540 men and 510 women).

The session-rate user attribute is *smaller* and *larger* which corresponds to whether or not a student-user's session-rate is less than or greater than the mean session-rate during each of the study-years (20.35 and 23.99 sessions per study-year respectively). There are 714/336 and 669/381 smaller/larger student-users during study-years one and two.

The conformance user-attribute is either *eclectic* or *conformant*. Being eclectic or conformant corresponds to whether or not the student-user undertakes a session during which only rare Websites are visited (is eclectic) or alternatively when locating Web information always visits one or more of the one thousand most visited Websites (is conformant). There are 395/655 and 294/756 conformant/eclectic student-users during study-years one and two respectively.

Changes in how student-users locate Web information during study-year two compared with during study-year one and which are described in the previous Section may be due to changes in (a) task, (b) Web structure or (c) the individual. An assumption of large scale real world Web information seeking is that there are no material changes in the users' information tasks. Individual change, in particular a *novice-effect* is examined in Chapter six.

Web structure including its information seeking affordances may be changing rapidly, hence any particular description of how student-users locate Web information may become quickly invalid. There is a sense in which such a description may ephemerally describe *the Web* at some point in time more than it describes *how users locate Web information*. Therefore the aim of the comparisons here is to identify descriptions which are consistent during each study-year, or time-invariant, and which are also consistent for different groups of student-users.[15] In this way a description of how student-users locate Web information is more reliable.

[15] This is a form of "triangulation", Denzin (1978).

For example, it might be tentatively concluded that student-users' session click rate during study-year two is greater than during study-year one. Overall this is the case as described previously. Each of the six separate attribute groups (men, women, smaller, larger, eclectic, conformant) also increase their session click rate as described below. Therefore the conclusion is more reliable. However, for comparison, one might conjecture that student-users Webhost-persistence increases. Overall this is also the case but the increase in women student-users Webhost-persistence is not significant ($z = 1.0$), hence the reliability of the overall conclusion that Webhost-persistence increases is weakened. The conclusion that, for example, men student-users' session click rate is greater than women student-users' must be correspondingly validated in each study-year (which it is). This validation by study-year provides a way of disambiguating some of the effects of structural change in the Web.

Similarities and differences between attribute group membership

User-attributes are not independent, for example student-users who have smaller session-rates are also more likely to be conformant. This enriches a description of how student-users locate Web information but it also distorts the assessment of reliability. The independence of each of the six combinations of pairs of user-attributes during both study-years is investigated using the χ^2 test for independence and the resultant[16] χ^2 test statistics are set out in Table 4.3. Four of the χ^2 test statistics exceed the critical value ($\chi^2_{1;0.005} = 6.63$) so these pairs of user-attributes, that is gender/session-rate and session-rate/conformance, are consistently (during each of the study-years) not independent ($p < .005$). The gender and conformance user-attributes are consistently independent ($p < .005$) of each other. That is, there is no association between a student-user's gender and whether or not the student-user never exclusively visits rare Websites when locating Web information.

[16] The associated underlying student-user frequency cross-tabulations are given in Table C.3 in Appendix C.

User-attribute combination	Study-year	
	study-year one	study-year two
gender/session-rate	$\chi^2 = 59.3$	$\chi^2 = 21.4$
gender/conformance	$\chi^2 = 0.43$	$\chi^2 = 0.001$
session-rate/conformance	$\chi^2 = 129$	$\chi^2 = 88.1$

Table 4.3: Cross-tabulation of χ^2 statistic for independence of pairs of user-attributes by study-year

Gender and session rate are associated with each other as are session-rate and conformance. Inspection of the student-user frequency cross-tabulations reveals that;

> *women* student-users are more likely (79% and 71% during study-years one and two respectively) to have a *smaller* session-rate than are men student-users (57% and 57%), and
>
> *smaller* session-rate student-users are more likely (48% and 38%) to be *conformant* than are larger session-rate student-users (13% and 11%)

On the other hand, gender and conformance are not associated with each other. Inspection of the student-user frequency cross-tabulations reveals that in this case;

> *women* student-users are as likely (37% and 28%) as *men* student-users (37% and 28%) to be conformant.

Similarities and differences between study-years one and two

Previously (see Table 4.2) an overall change between study-years one and two was found in each of the seven user-characterization metrics. This overall conclusion regarding change between during study-years one and two is now tested within each of three user-attribute groups. The change in how student-users locate Web information is validated separately in each of three user-attribute groups for three out of the seven user-characterizations. These three are;

1. average session click rate

2. average query-click proportion, and

3. average Website-re-request rate.

The z test statistic of the differences[17] (between study-years one and two) in respect of these three user-characterization for each of the six user-attribute groups is shown in Table 4.4. All the z test statistics of difference exceed the critical value $z_{0.01} = 2.33$ so that, for example, men student-users' average session click rate during study-year two is greater than that during study-year one ($p < .01,\ z = 7.4$). Thus the reliability of the conclusion that during study-year two all of these user-characterizations increased is improved because the conclusion is valid separately for each of the six student-user attribute groups.

User-attribute partition	**User-characterization**		
	average session click rate	average query-click proportion	average Website-re-request rate
men	$z = 7.4$	$z = 8.5$	$z = 2.8$
women	$z = 9.4$	$z = 10.2$	$z = 7.5$
small	$z = 10.3$	$z = 12.5$	$z = 9.9$
large	$z = 4.5$	$z = 9.9$	$z = 3.3$
conformant	$z = 7.8$	$z = 6.0$	$z = 6.5$
eclectic	$z = 7.9$	$z = 14.1$	$z = 7.1$

Table 4.4: Cross-tabulation of z statistic for study-year difference by user-attribute partition for three user-characterizations

A possible interpretation is therefore that these three user-characterizations are reflecting structural changes in the Web. For example, structural changes in the way that Web pages are constructed or in the way that information is distributed over the Websites within a Webhost could affect the information seeking affordances of the Web. The change in session click rate may also be as a result of a general lengthening of session duration. In Chapter five the comparison of search-queries between during study-years one and two suggest that session durations remain the same which strengthens the suggestion that it is Web information seeking affordances which have changed, and in particular that using the Web during study-year two demands more query-clicking.

The evidence of a change between study-years one and two for the other user-characterizations which include both of the session-to-session user-characterizations

[17] The underlying user-characterizations by user-attribute are given in Tables C.4 and C.5 in Appendix C.

is mixed. These four user-characterizations are;

1. average Webhost-persistence
2. average session-conformance
3. average Website-trajectory slope, and
4. session-conformance range.

The absence of change is most evident in the group of larger (than average session-rate) student-users who show no significant difference in any of these user-characterizations.

The overall change for these four user-characterizations is focussed in a particular group of student-users where the change is pronounced and sufficient to give the results discovered earlier. For example, men student-users increase their average Webhost-persistence from 2.25(0.03) Websites per Webhost to 2.37(0.03) Websites per Webhost ($p < .01,\ z = 2.8$) but the difference in average Webhost-persistence for women is not significant ($z = 1.0$). (Despite this, there is no gender difference for average Webhost-persistence although the numerical relation

$$\text{user-characterization}_{\text{men}} < \text{user-characterization}_{\text{women}}$$

during study-year one reverses during study-year two.)

The overall reduction in average session-conformance is most pronounced in the group of smaller student-users who reduce ($p < .01,\ z = 5.6$) from 1.16(0.007) to 1.10(0.008) while the group of larger student-users remain at 1.12 average session-conformance. (As with the average Webhost-persistence user-characterization, the numerical relationship between the user-attribute groups reverses from one study-year to the next.)

The flattening in the Website-trajectory slopes is more pronounced among women student-users for whom the slope reduces ($p < .01,\ z = 6.4$) from 0.42(0.006) Websites per click to 0.37(0.005). There is again a reversal of the numerical relationship. Although there is a gender difference during study-year one ($p < .01,\ z = 3.8$) this disappears during study-year two ($z = 1.6$). There is also a pronounced flattening of Website-trajectory slope among the group of smaller student-users but this may be due to the association between the gender and session-rate user-attributes among student-users, especially since the group of larger student-users' Website-trajectory slopes remains the same at 0.35 Websites per click.

Overall the session-conformance range increases. There is a pronounced increase ($p < .01,\ z = 5.7$) among the group of smaller student-users (from 0.90(0.02) to

1.06(0.02)) and also ($p < .01$, $z = 4.2$) among the women student-users (from 1.04(0.03) to 1.19(0.02). Table 4.5 sets out the z test statistic of difference for each user-attribute partition of student-users in respect of the four user-characterizations just discussed.

User-attribute partition	User-characterization			
	average Webhost-persistence	average session-conformance	Website-trajectory slope	session-conformance range
men	$z = 2.8$	$z = 4.2$	$z = 1.6$	$z = 3.1$
women	$z = 1.0$	$z = 2.1$	$z = 6.4$	$z = 4.2$
small	$z = 2.1$	$z = 5.6$	$z = 6.2$	$z = 5.7$
large	$z = 1.2$	$z = 0.0$	$z = 0.0$	$z = 0.7$
conformant	$z = 0.9$	$z = 2.0$	$z = 4.7$	$z = 2.8$
eclectic	$z = 2.5$	$z = 2.0$	$z = 3.1$	$z = 2.8$

Table 4.5: Cross-tabulation of z statistic for study-year difference by user-attribute partition for four user-characterizations

Similarities and differences within attribute groups

A tentative conclusion of a difference within an attribute group, say a gender difference, is validated by this being the case during each of the two study-years. This is described as a *consistent* difference so that a consistent difference for a user-attribute group is where the difference within the group in respect of a user-characterization applies during each of the two study-years (and the difference is in the same direction). Hence for example, the gender user-attribute group shows consistency as regards average session click rate since during study-year one the characterization for men is 32.4(1.42) but for women is 25.3(0.86) ($p < .001,\ z = 4.3$) while during study-year two characterization for men is 50.0(1.89) but for women it is 37.9(1.02) ($p < .001,\ z = 5.6$). (That is there is a significant difference between men and women in respect of average session click rate during each of study-years one and two with the men's average session click rate being greater than that of the women.)

Table 4.6 illustrates the user-attribute/user-characterization combinations where there is a consistent significant difference within the user-attribute group. Only five of the user-characterizations are illustrated because the average Website-re-request rate and average Webhost-persistence user-characterizations do not show any consistent differences. Table 4.6 sets out only the z test of difference statistic for each study-year where this is significant ($\alpha = 0.05$). The conformance attribute groups of student-users show a large consistent difference for both the conformance based user characterizations (average session-conformances and session-conformance range) by definition. Elsewhere in the Table the attribute group membership associations between gender and session-rate, and between session-rate and conformance may explain the gender difference in average session click rate and conformance difference in average query-click proportion.

User-characterization	User-attribute		
	gender	session-rate	conformance
average session click rate	$z = 4.3$ and $z = 5.6$ (men > women)	$z = 4.2$ and $z = 2.8$ (smaller < larger)	
average query-click proportion		$z = 3.1$ and $z = 4.24$ (smaller > larger)	$z = 5.6$ and $z = 2.24$ (conformant > eclectic)
average session-conformances			$z = 26.0$ and $z = 22.6$
Website-trajectory slope		$z = 11.3$ and $z = 7.1$ (smaller > larger)	
session-conformance range		$z = 17.3$ and $z = 12.4$ (smaller < larger)	$z = 110$ and $z = 110$

Table 4.6: Cross-tabulation of z statistics for consistent user-attribute partition difference by user-characterization and user-attribute

Table 4.6 can be used to disambiguate some of the effects of structural change which is possibly causing a change in average session click rate, average query-click proportion (and average Website-re-request rate).

The change in average Website-re-request rate appears to apply uniformly to all of the user-attribute subgroups and hence the suggestion that this change is structural is strengthened. The increase in average session click rate is not uniform, thus while there may be structural influences, there appears also to be a gender/session-rate influence. Similarly the average query-click proportion may be structural but is also influenced by session-rate/conformance.

Table 4.6 demonstrates that there is no general gender difference as regards the user characterization metrics. The only possible candidate, the difference in average session click rate, can be alternatively explained by the association between the gender and session rate attribute. That is, the difference in average session click rate is a feature primarily of session rate attribute not gender attribute.

Similarly the association between the session rate attribute and the conformance attribute provides an explanation for the difference in the average query-click proportion within the conformance attribute group.

Thus the simplest interpretation is that there are just four reliable differences in the user-characterizations all of which can be associated with session rate. These are that the;

average session click rate is less for student-users who have smaller session rates (and greater for men),

average query-click proportion is greater for student-users who have smaller session rates (and greater for conformant student-users),

Website-trajectory slope is greater for student-users who have smaller session rates, and

session-conformance range is smaller for student-users who have smaller session rates.

The conclusion of the analyses of similarities and differences in how student-users locate Web information undertaken here can be summarised as;

1. there is no gender difference in the user characterizations,
2. there is a structural increase in average Website-re-request rate,

3. some of the increase in average session click rate and average query-click proportion is structural,

4. there are four differences which are associated with a student-user's session rate, namely;

   average session click rate,

   average query-click proportion,

   Website-trajectory slope, and

   session-conformance range.

   The user characterizations of student-users with larger session rates are time invariant.

5. Webhost-persistence is both time invariant and similar among all the attribute groups of student-users, and lastly,

6. conformant and smaller student-users are associated and are more active, that is have a higher average query-click proportion than do eclectic/larger student-users.

In the next Section, the Web log analyses consider the diversity of Websites which student-users visit session-to-session.

4.4 Website popularity

The previous meta-analyses of the Web log describe how student-users locate Web information in terms of numerically based user-characterizations such as Website re-request rate. What Website it is that is revisited is not relevant (and not revealed). Even the session-conformance based user-characterizations consider only whether the same collection of Websites are bring visited during each session so that particular Websites are not relevant. In this Section the investigation reveals which particular Websites student-users visit in order to locate Web information. The analysis is session-to-session since it considers student-users' study-year vocabulary.

Web 'popularity' can be determined in at least four different ways. Each way purports to facilitate a ranking which compares one Web service with another Web service (a Web service may range from being a particular Website to a collection of Webhosts). Achieving a high ranking position, that is being more 'popular', is important for many commercial Web services since it can determine the price charged for Web advertising. 'Popularity' metrics are based on,

1. 'hit rates' or the number of 'visits' made to a Web service,

2. 'session' counts or the number of 'sessions' during which a Web service is visited regardless of how many times within a 'session' it is visited,

3. 'user' counts or the number of 'individuals' who 'visit' the site regardless of the number of 'sessions' in which they participate, and,

4. 'visit' duration that is the length of time that 'users' remain at a Web service,

but there is no consensus on the definition of the terminology which is employed.

Hit rate and visit duration based metrics are both only noted here for completeness. Hit rate data is generally discredited as a Web metric (for example Pitkow, 1997) and reliable duration data cannot be obtained from transaction logs (either server-side or client-side). These issues are discussed in Chapter three.

The session-conformance metric in the previous Section uses the session frequency of a Website during both study-years (that is the number of sessions during which Web requests are made to a particular Website) to construct a ranking of Websites and attention is focussed on the top one thousand such Websites. Hence this investigation into how student-users locate Web information seeking is already informed by a 'session popularity'. From this, it is clear that many student-users visit a few Websites and many Websites are visited by few student-users. At the extreme, 83,866 Websites that is over 83% of the Website-repertoire during both study-years have a session frequency of only one and therefore were visited by just one student-user.

Intuitively a *popularity* metric should correspond to a count of *people* so that popularity information is not immediately accessible from session frequencies since if, for example, a small number of student-users with high session-rates visit Websites not generally visited, then, by-session, these Websites will appear 'popular' notwithstanding that only a few student-users ever visit them.

The *individual-popularity* of a Website is thus defined to be the frequency of student-users who include the Website in their vocabulary. Hence the individual-popularity of a Website during a study-year is the frequency of student-users who include the Website in their study-year vocabulary.

The *relative-individual-popularity* metric for a Website is the *proportion* of user-students whose vocabulary includes the Website, that is,

$$\text{Website relative-individual-poularity} = \frac{\text{Website individual-popularity}}{\sum \text{student-users}}$$

$$= \frac{1}{1,050} \text{Website individual-popularity.}$$

The distribution of Website individual-popularity during study-year two[18] is illustrated in Figure 4.17. The most popular Website during study-year two is visited by 768 (or 73%) of the 1,050 student-users; individual-popularity does not become minimal (= 1 student-user) until rank 3,867. For study-year one these values are 67% and 2,255. Hence during study-year two, while the most popular Website has become more individually-popular, popularity > 1 has also become more diverse and includes over 71% more (from 2,254 to 3,866) Websites. This implies that the 'concentration' of popularity in Websites changes over time and suggests that different groups of student-users may have different popularity distributions.

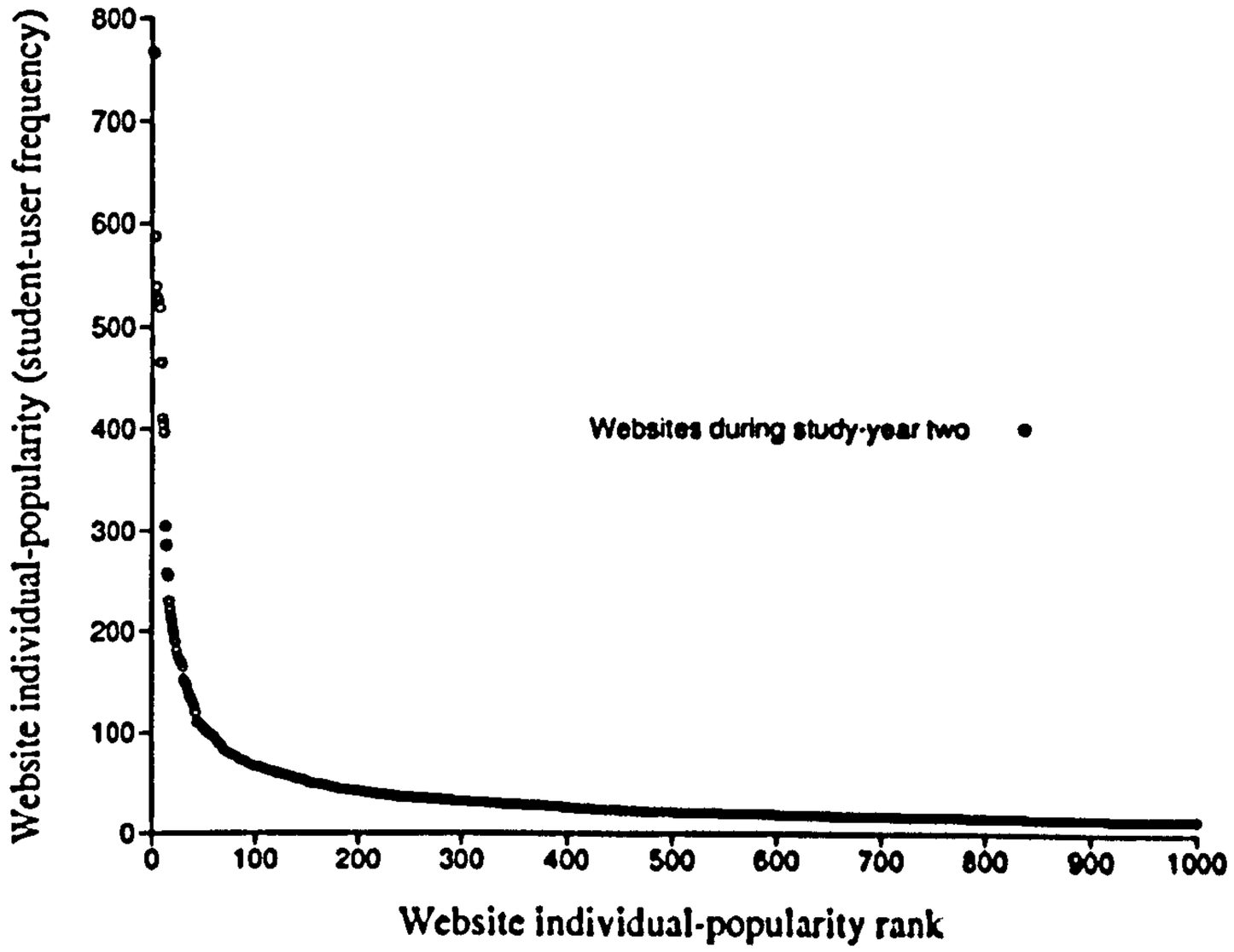

Figure 4.17: Zipf distribution of Website individual-popularity during study-year two

The relative-individual-popularity of the 'top-twenty' ranked Websites during study-year two[19] is shown in Table 4.7. The most popular Websites illustrate the dominance of major Web brands such as Yahoo! which provides six out of the seven Websites which occur in the 'top-twenty' during each study-year. On the other hand most of the Websites in the study-year one top-twenty do not appear in the study-year two top-twenty. For example, during study-year one Netscape provides nine Websites in the top-twenty but this reduced to three Websites during study-year two. This

[18] The equivalent graph for study-year one is illustrated in Figure C.14 in Appendix C.
[19] See Table C.6 in Appendix C for study-year one.

reinforces the interpretation that how student-users locate Web information changes over time but this may be a consequence of structural change in the Web, for example by consolidation[20] or disaggregation of Websites. All the Websites shown relate to the provision of Web search services. How student-users use search services (both search engines and directories) is discussed in Chapter five.

Rank	Proportion of student-users	Website (conditioned url-string)
1	73.1%	<utility3-search.europe.yahoo.com/search/ukie>
2	73.0%	<home.europe.yahoo.com/>
3	56.0%	<google.yahoo.akadns.net/bin/query_uk>
4	51.3%	<homerc.europe.yahoo.com/>
5	50.4%	<altavista.com/cgi-bin/query>
6	50.1%	<www.altavista.magallanes.net/cgi-bin/query>
7	49.4%	<altavista.com/cgi-bin/query>
8	44.4%	<www.yahoo.akadns.net/>
9	39.1%	<www.infoseek.com/Home>
10	38.6%	<www.infoseek.com/Titles>
11	37.8%	<search.snv.yahoo.com/bin/search>
12	29.0%	<search.yahoo.co.uk/search/ukie>
13	27.2%	<google.yahoo.com/bin/query>
14	24.6%	<cgi.netscape.com/cgi-bin/plug-in_finder.cgi>
15	24.4%	<search.snv.yahoo.com/search>
16	21.9%	<netscape.google.com/netscape>
17	21.1%	<www.savvysearch.com/>
18	20.5%	<members.tripod.com/adm/popup/roadmap.shtml>
19	20.0%	<www.altavista.magallanes.net/av/eng/help.htm>
20	19.2%	<search.netscape.com/cgi-bin/search>

Table 4.7: Relative-individual-popularity of 'top-twenty' Websites during study-year two

Table 4.7 also shows that the twentieth most popular Website during study-year two attracts 19.2% of student-users. During study-year one the equivalent proportion is 30.1%. This suggests that during study-year two, student-users visit a more diverse collection of Websites compared to study-year one which conclusion is compatible with that in the previous Section which concluded that during study-year two, how student-user locate Web information is more eclectic compared to during study-year one.

However interpretation of the individual-popularity metric requires care since the individual-popularity of a particular Website may be associated with the individual-popularity of another Website. For example the 73.1% of student-users who visit the

[20] The url-string conditioning procedure already consolidates the Netscape Websites <www-uk.netscape.com/uk/escapes/search/ntsrchrnd-*.html>.

top ranked Website might include all of those student-users who also visit the second ranked Website. In consequence the total frequency of different student-users who visit a collection of Websites may not increase with the individual-popularity rank of Website.

The *collective-popularity* of a collection of two or more Websites is defined to be the frequency of student-users who include at least one of the Websites from the collection in their Web vocabulary. The corresponding relative-collective-popularity metric (of a collection of Websites) is therefore the proportion of student-users who include at least one of the Websites from the collection in their Web vocabulary. This metric is now used to compare the 'concentrations' of popularity by determining how many Websites need to be in the collection[21] in order that 95% of all student-users should have visited at least one of the Websites from the collection. This is equivalent to 5% of student-users not visiting *any* of the Websites in the collection.

Figure 4.18 illustrates the distribution of the relative-collective-popularity of the top-twenty Websites ranked by individual-popularity for each study-years As expected (given the apparently less disperse popularity distribution of study-year one) the relative-collective-popularity distribution during study-year one appears to be more concentrated and reaches 95% before the distribution for study-year two. This is despite the top ranked Website during study year one being less individually-popular than the top-ranked Website during study-year two.

[21] That is, in the sense of how far down the ranking order one must descend.

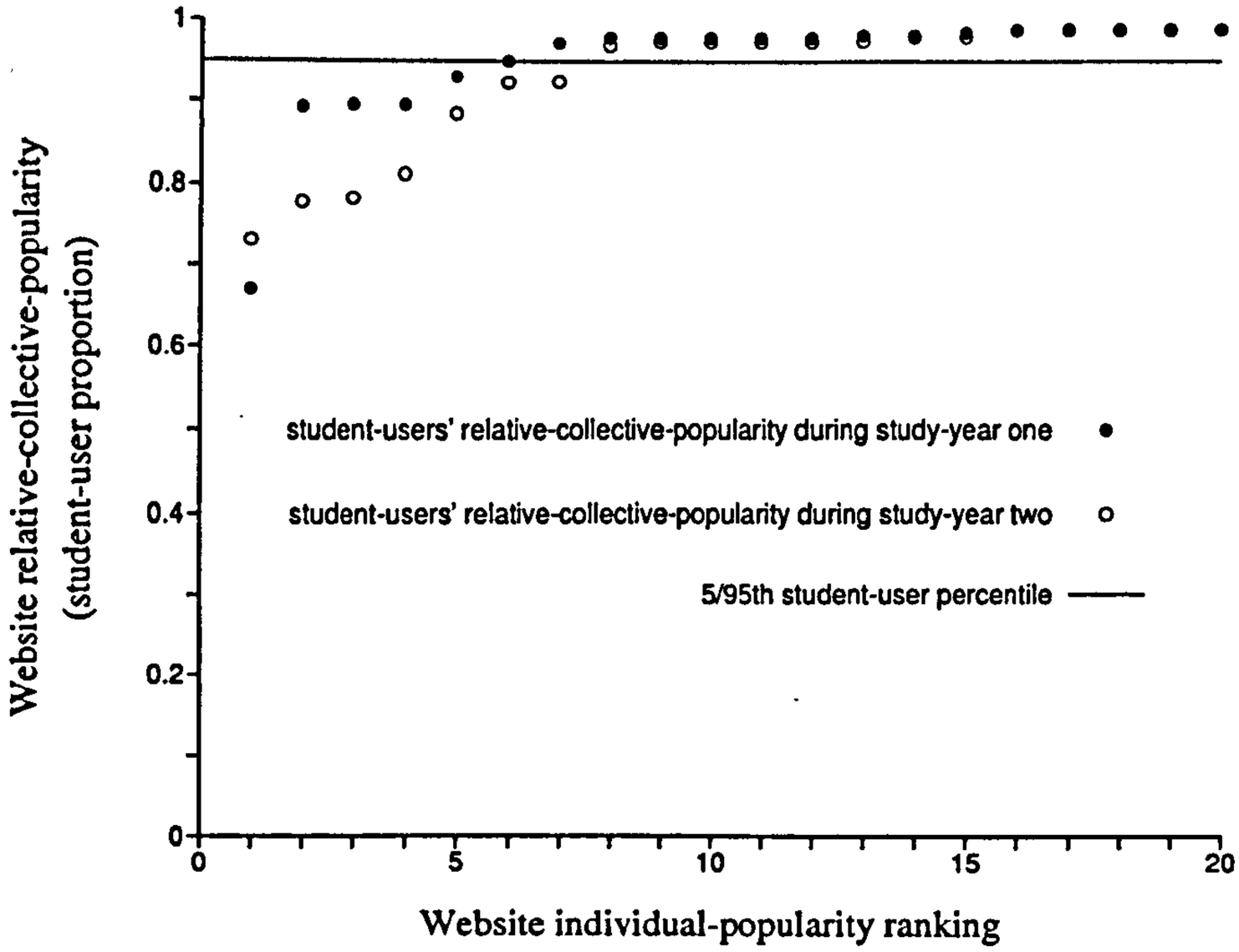

Figure 4.18: Ranked distributions of Website relative-collective-popularity by Website individual-popularity

Figure 4.18 therefore adds weight to the interpretation that student-users are more diverse in how they locate Web information during study-year two compared to during study-year one but this is by a lesser extent than might be inferred from the 19.2%:30:1% ratio of twentieth rank Website relative-individual-popularity.

At rank twenty, the collective-popularity is 1,039 student-users and 1,038 student-users for study-years one and two respectively. Table 4.8 gives the collective-popularity for study-years one and two for the top-twenty ranked Websites during each study-year. This illustrates the phenomenon of association between Websites which appears soon in terms of ranking position (at ranks three and four during study-year one) and which becomes very strong after the 95% ($\approx$ 998 student-users) level of relative-collective-popularity is achieved. 100% relative-collective-popularity (that is every student-user has visited at least one of the Websites in the collection) occurs at rank 76 during study-year two but perversely (since otherwise popularity is more concentrated) not until rank 105 during study-year one.

Rank	Collective-popularity	
	during study-year one (student-users)	during study-year two (student-users)
1	705	768
2	938	817
3	941	822
4	941	853
5	977	930
6	997	969
7	1,020	970
8	1,027	1,017
9	1,027	1,022
10	1,027	1,022
11	1,027	1,022
12	1,027	1,022
13	1,030	1,023
14	1,030	1,028
15	1,034	1,028
16	1,038	1,037
17	1,039	1,037
18	1,039	1,037
19	1,039	1,037
20	1,039	1,038

Table 4.8: Collective-popularity of 'top-twenty' Websites by study-year

The similarity between the study-years in collective-popularity in Table 4.8 which contrasts with the differences in individual-popularity is explained by student-users' Web information seeking during study-year one being more homogeneous in the sense that visits to the 'top-twenty' Websites are spread more evenly across student-users. During study-year two more student-users eschew particular more individually-popular Websites which adds weight to the interpretation that there is greater individual diversity of Web information seeking by student-users during study-year two compared with during study-year one.

Figures 4.19 and C.16 in Appendix C illustrate the differences in collective-popularity during each of the study-years that result from student-users being either eclectic or conformant. Eclectic Web information seekers undertake sessions which are exclusively visits to rare Websites (that is Websites with a session frequency < 1000) while conformant Web information seekers always include visits to non-rare Websites in their information seeking sessions. The graphs show that eclectic student-users are more likely to visit at least one of the more individually-popular Websites than are

conformant student-users. Hence the sessions of visits to rare Websites by eclectic Web information seekers *complement* rather than substitute for their visiting the more-individually popular Websites. Reciprocally, conformant student-users do not concentrate their Web information seeking just in the most popular Websites; they just always include non-rare Websites in how they locate their Web information.

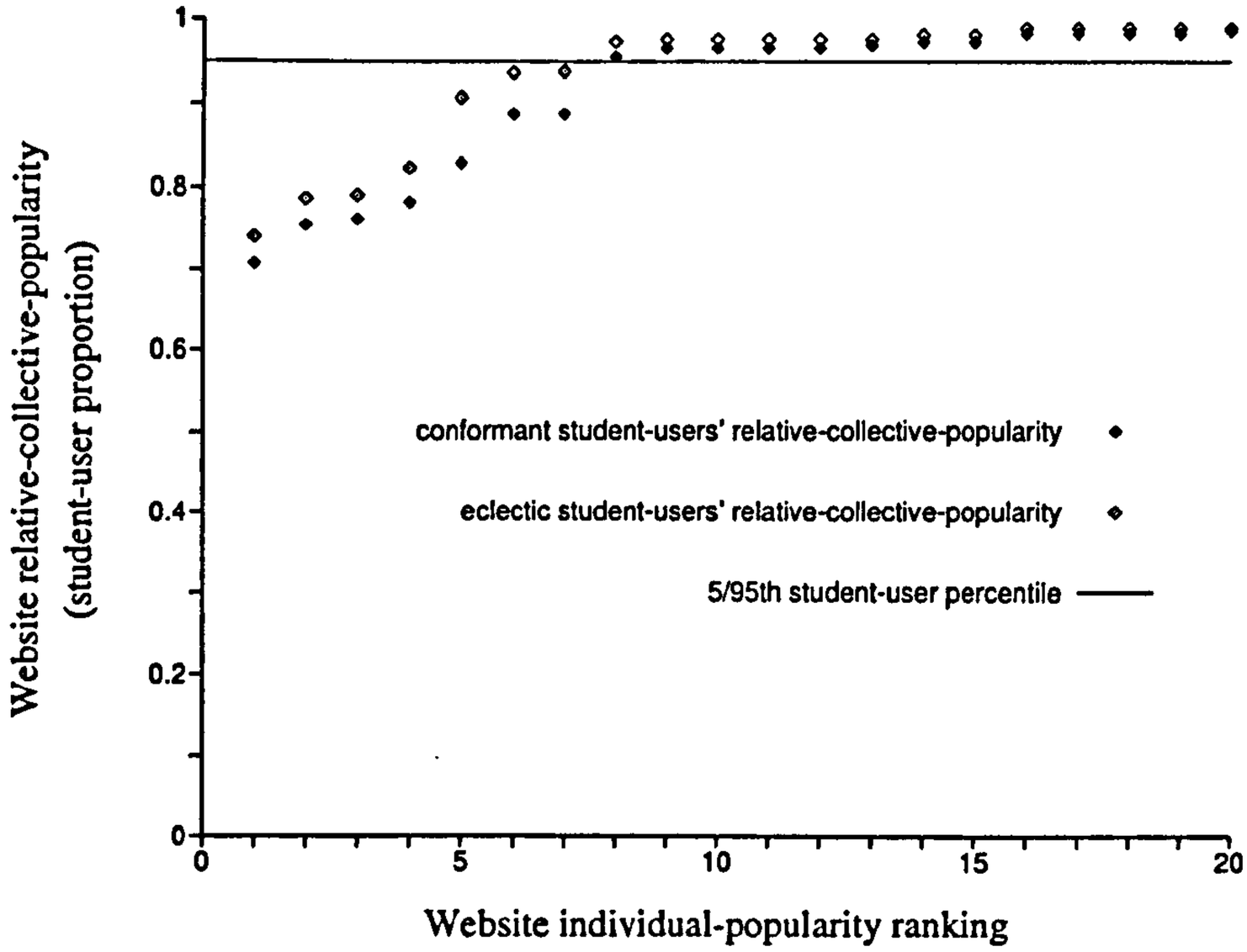

Figure 4.19: Ranked distributions of Website relative-collective-popularity by Website individual-popularity for conformant and eclectic student-users during study-year two

Collective-popularity is sensitive to including or excluding Websites from the collection under consideration. In the extreme, if a default 'home' page were included then its 100% individual-popularity would vitiate the analysis. The Websites (or conditioned url-strings) which are considered in this analysis exclude some Websites, see Chapter three, otherwise this possibility may have been encountered.

Figures 4.20 and 4.21 provide a form of sensitivity analysis for the collective-popularity metric. From the comparison between the two study-years in Figure 4.18 it was concluded that Web information seeking was less diverse during study-year one than during study-year two because 95% of student-users during study-year one visited at least one Website out of a *smaller* collection of Websites than during study-year two. Comparison of Figures 4.20 and 4.21 shows that this argument remains valid as the most individually-popular Websites is progressively excluded from the collective-popularity computation. For example, if the first three most individually-popular Websites are excluded so that the collection starts at rank four then, during study-

year one, nine Websites are needed in the collection before the relative-collective-popularity exceeds 95% while during study-year two thirteen Websites are needed.

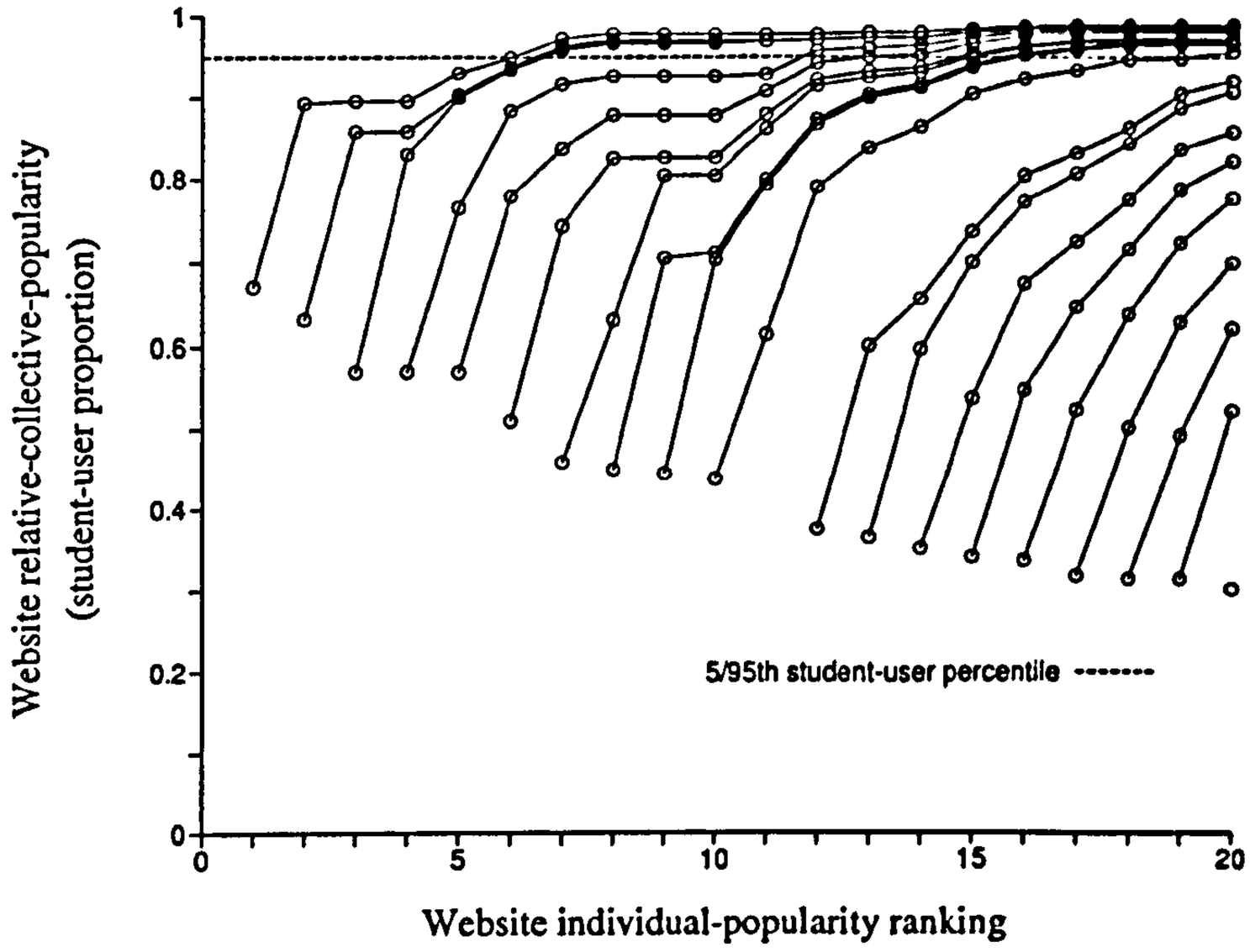

Figure 4.20: Website collective-popularity distribution family during study-year one

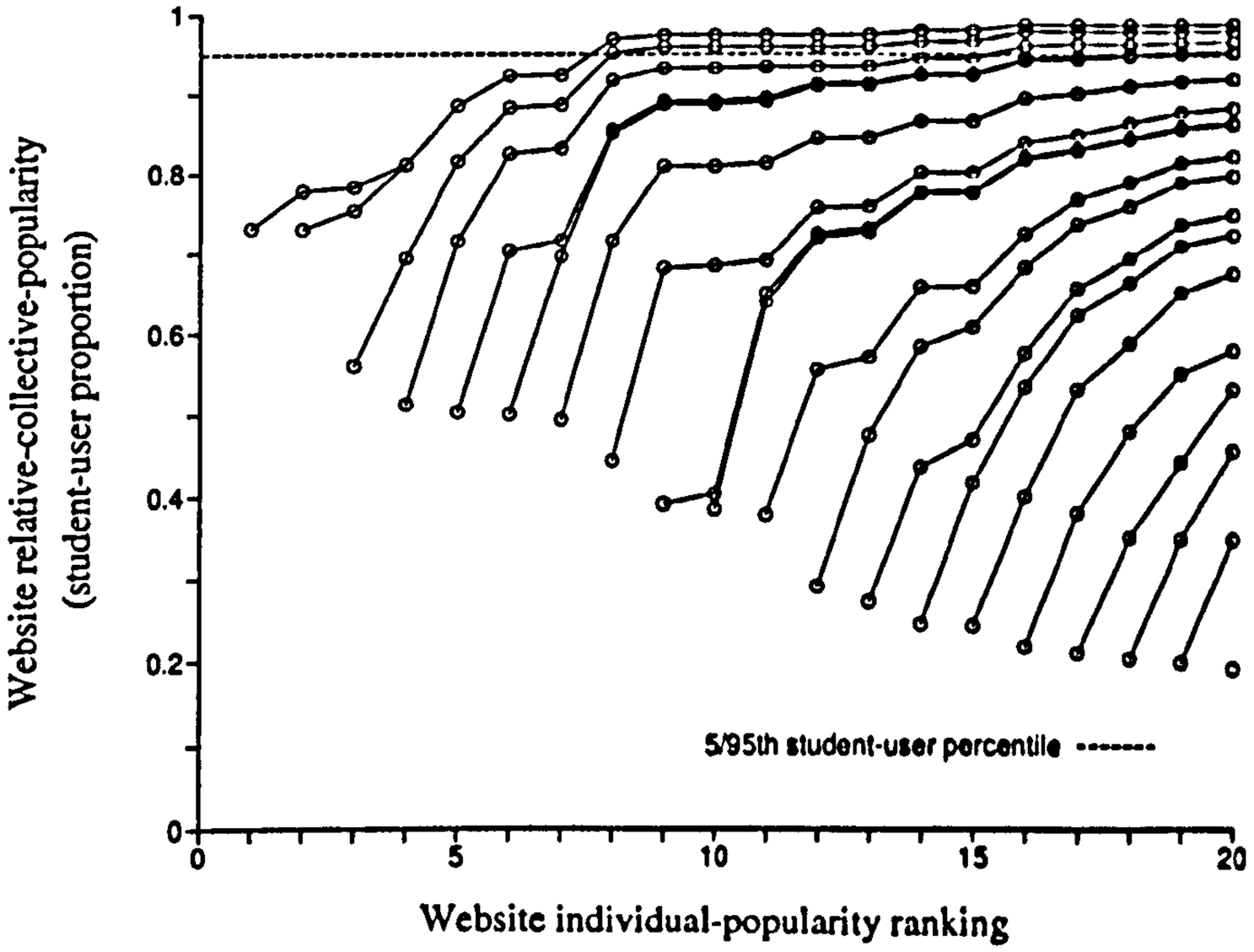

Figure 4.21: Website collective-popularity distribution family during study-year two

The relative-collective-popularity metric is therefore sufficiently robust to claim that this session-to-session analysis of study-year vocabulary shows that;

1. during study-year two, student-users locate Web information by visiting a more diverse collection of Websites compared with study-year one and rely less on visits to the more individually-popular Websites, and

2. eclectic student-users visit the more individually-popular Websites more than do conformant student-users so that visiting rare Websites exclusively does not substitute for visiting individually-popular Websites, rather how eclectic student-users locate Web information is an expansion of how conformant student-users locate Web information in the sense that eclectic student-users *in addition* undertake sessions which are exclusively visits to rare Websites compared to conformant student-users whose Web information seeking always includes visits to non-rare Websites.

4.5 Summary and discussion

Overall conclusions about how student-users locate Web information are determined by using seven user-characterizations to interpret the Web log. These are (i) average session click rate, (ii) average query-click proportion, (iii) average Website-re-request rate, (iv) average Webhost-persistence, (v) Website-trajectory slope, (vi) average session-conformance, and (vii) session-conformance range. The conclusions are refined by considering the similarities and differences between and within groups of student-users. These groups are defined according to three user-attributes (gender, session-rate and conformance). The analyses in this Chapter also consider Website popularity in order to investigate diversity in session-to-session Website visiting.

The Website-trajectory slope and session-conformance range user-characterizations are session-to-session. They thus characterize in an extended or macro sense (across sessions) how a student-user locates Web information while the other user-characterizations describe session-by-session how a student-user locates Web information in a more micro sense within a session. How student-users locate Web information session-to-session resembles how student-users locate Web information session-by-session so that the extended Web information seeking of a student-user across subsequent sessions is the same as his (or her) micro Web information seeking within each session separately.

Overall, during study-year one, each student-user typically locates Web information by employing,

$$29.0 \text{ (clicks per session)} \approx 28.06$$
$$= 2.0 \text{ (clicks per Website)} \times 2.3 \text{ (Websites per Webhost)} \times 6.1 \text{ (Webhosts per session)}$$

while, during study-year two, this 'session-equation' becomes,

$$44.1 \text{ (clicks per session)} \approx 45.264$$
$$= 2.3 \text{ (clicks per Website)} \times 2.4 \text{ (Websites per Webhost)} \times 8.2 \text{ (Webhosts per session)}.$$

When study-year two is compared with study-year one, how student-users locate Web information is more *energetic* or entails more clicking, and is more *active* since there is a greater proportion of query-clicking. Compared with study-year one, during study-year two student-users' session magnitude increases, that is their Website and Webhost session repertoires are greater. Student-users also revisit more Websites and visit more different Websites within each Webhost during study-year two compared with study-year one.

In particular, during study-year two student-users' average Web information seeking is more similar than it is during study-year one. But student-users during study-year two are also more eclectic and individuals become more distinctive in how they locate Web information.

The finding that during study-year two student-users each visit a more diverse collection of Websites than during study-year one corroborates Barford *et al.* (1999, p. 22) who concluded that "relatively speaking, the most popular [Websites] are less popular in 1998 ... than in 1995" (see Chapter two).

On the face of it these findings are difficult to interpret. The usual model of task specificity correlates specificity inversely with the number of clicks. This suggests that during study-year two users' information tasks become much less specific. However it is assumed that as a result of the large scale real world context of the investigation, task specificity is unchanged. In consequence the findings may be interpreted as suggesting that users become much *less* systematic or browse *more* as they become more experienced which directly opposes Marchionini's searching/browsing thesis.

An alternative form of explanation is that session duration and/or Web structure has changed. At the click level, the findings are therefore ambiguous.

There is considerable variation among student-users' user-characterizations. For most student-users the mean user-characterization overstates how they locate Web information because of the asymmetry of distribution which results from the underlying power law distribution of session-rate and session click rate metrics.

At a more macro level of Web information seeking the key initial finding is that,

> student-users can be categorised into two groups by how they locate Web information;
>
> (a) *conformant* student-users whose Web information seeking always includes visits to the more frequently visited Websites, and
>
> (b) *eclectic* student-users whose Web information seeking sessions sometimes comprise exclusively visits to infrequently visited (or rare) Websites. These sessions supplement eclectic student-users' Web information seeking in that the visiting of rare Websites is in addition to visiting the more frequently visited Websites.

An interpretation of this information seeking behaviour is that eclectic student-users are building/have built for themselves individually distinctive Web *territories* or collections of Website requests which satisfy their personal information needs. Conformant student-users, on the other hand may be at an earlier stage of building their

territories so they are not as individually distinctive. The idea of territory here is more than just the Website vocabulary since it entails also the proportions of clicks made to each Website during a session.

Marchionini (1995, p. 11) constructs the notion of a "*personal information infrastructure*" in order to describe "an individual person's collection of abilities, experience, and resources to gather, use, and communicate information". He continues to explain that the "level of development of a person's information infrastructure is roughly analogous to the level of his or her information literacy". Marchionini understands information seeking in electronic environments as several different electronic environments. Thus, for him a person's information infrastructure includes knowledge (for example its information organisation) specific to one or more particular electronic environments.

The notion of a personal *Web* information infrastructure is useful here to understand Web 'territories'. What is a personal Web information infrastructure follows Marchionini's description except that it relates to Web information and therefore applies to just a single electronic environment, the Web. As with his original notion it is hypothesized here that "experience with a variety of [Web] information problems ... leads us to develop both a general knowledge of how [Web] information is organized and the skills needed for facilitating access to it". In addition "as we gain experience with [Web] information problems, we strengthen our general [Web] information-seeking knowledge and skills" (Marchionini, 1995, p. 13).

The comparison of similarities and differences within student-user attribute groups and over time disambiguates some of the effects of structural change in the Web and facilitates a refinement of the overall descriptions of how student-users locate Web information.

It is concluded that;

> Men and women student-users are the same in respect of how they locate Web information.
>
> Student-users are all similar in their Webhost-persistence.
>
> Conformant student-users have smaller session rates and are more active than eclectic student-users. That is, conformant student-users use querying rather than passive link-clicking more than do eclectic student-users.
>
> Student-users with larger session rates have flatter Website-trajectory functions hence their information seeking appears to be more focussed in already visited Websites when compared to student-users with smaller session rates.

> Student-users with larger session rates are more consistent in how they locate Web information, that is, their information seeking during study-year two and during study-year one is more similar than the information seeking of student-users with smaller joint session rates.

These findings support the notion of student-users' territory building. For example, conformant student-users rely more on querying which suggests that compared to eclectic student-users they are less sure of where Web information which may satisfy their niche information needs is located. Eclectic student-users by comparison might be said to know where they are going or how to satisfy their niche information needs. In addition, those student-users with larger session rates might be expected to be more advanced in their territory building and, having determined the Websites which satisfy their particular niche information needs, be more similar in how they locate Web information when comparing study-year two with study-year two. This is found to be the case.

However, in this Chapter, the argument can be applied only to groups of student-users. In order to make proper use of the notion of a personal Web information infrastructure then this must be applied at the level of the individual.

5

How do student-users use Web information location services?

5.1 Introduction

Do all student-users use Web information location services and do they all use such services to the same extent?

The term Web *search-engine* is commonly used to refer generically to Web search services although distinctions are sometimes drawn between 'pure search-engines' and services such as Yahoo! which rely on human cataloguing of Web resources. Meta search-engines such as Metacrawler which search search-engines also form a distinct service.

All Web search services have three constituent components which (a) identify/collect Web resources, (b) index/catalogue the resource, and (c) facilitate retrieval of those resources relevant to a user enquiry. The original defining characteristic of a Web search-engine was the use of robot software which crawled the Web and automatically reported Website indexing information to the search-engine. This distinguishes only the first constituent component of the service (and which is transparent to the user).

This Chapter is based on distinguishing services by the third component, that is the mechanism by which the student-user interacts with the search service. There are two possibilities. The user can either submit a 'search-query' (in the information retrieval sense) or follow a hierarchical hypertext topic link. Whether or not a search service has used a robot does not predicate which of these interaction mechanisms the student-user chooses to use. The two forms of interaction, *query-clicking* and *link-clicking* can be reliably distinguished since each query click contains a search-part in the url-string (see Chapter three).

Web search-queries and *Web information location services* are defined recursively since,

Web search-queries are search-queries submitted to a Web information location service Website, and,

Web information location services provides Websites to which Web search-queries can be submitted.

In addition Web search-queries are associated with a particular form of search-part (see Chapter three) so that 238,614 Web search-queries and fourteen Web information location services can be identified by analysing the Web log. This is described below.

For brevity, a Web search-query from now on is referred to as just *search-query*. A *search-session* is a session which includes a search-query and a *search-user* is a student-user who has a search-session.

As in Chapter four, the investigation in this Chapter uses the gender, session-rate and conformance attributes (men/women, smaller/larger, and conformant/eclectic) to classify student-users and facilitate interpretation of the Web log. The analyses in this Chapter also use four characterization metrics specifically constructed to describe and differentiate how student-users use Web information location services. These are;

1. average search-query proportion,
2. search-session proportion,
3. average search-query count, and
4. average search-term count,

which are defined below.

The analyses in Section 5.2 address the conjecture that every student-user uses Web information location services. Although this is shown to be true, the analyses reveal a wide variation in the likelihood that any individual student-user will use Web information location services.

In Section 5.3 there is an analysis of search-queries submitted to the AltaVista and Excite Web information location services (the *AltaVista-Excite sample*). This includes a comparative analysis between the findings from the AltaVista-Excite sample and the published surveys of studies of Excite search-queries. Most search-queries

are composed of only one or two terms. *Singleton* search-queries are search-queries which contain a single search-term only. Search-users who have a smaller session rate are more likely to use singleton search-queries but the likelihood of using singleton search-queries is independent of the search-user's gender. Some search-users use singleton search-queries exclusively.

Section 5.4 provides a summary of the Chapter and includes a discussion of the results of the analyses.

5.2 Web information location service usage

In Chapter four, query-clicks are defined as Web request clicks which contain a search-part and query-sessions are defined as sessions which contain a query-click. The occurrence of querying, that is query-clicks and query-sessions, is examined and it is found that, overall, both increase during study-year two compared to during study-year one. However the inclusion by Websites of search-parts in the url-string is a widespread technique and does not of itself indicate a search-query. During the two study-years of the investigation, query-clicks are submitted to 9,793 different Webhosts. As well as Web information location services these include all the other Websites visited which request user data.

A simple popularity analysis of the Web log can find the principal Websites to which query-clicks are submitted and hence the principal Web information location services which have been used can be identified by analysing the search-part. This procedure is described in Chapter four[1] and results in identifying fourteen principal Web information location services.

A Web information location service is associated with one or more Websites which use an interface based on the information retrieval *search-query* paradigm, that is, the user submits a search-query consisting of one or more *search-terms* and the (information retrieval) system responds. Some Web services, most notably Yahoo! also have a hierarchical directory interface as well as a Website using a search-query interface. It is only information seeking activity by means of the search-query interface which is considered here. A *search-query-click* is defined as a query-click to one of these fourteen principal Web information location services (which are listed below). For the sake of brevity the qualification 'principal' is now understood.

The overall proportion of query-clicks (or $\frac{\sum \text{query-click rate}}{\sum \text{click rate}}$, query-clicks per click) during each of the study-years has already been seen to have increased from 26% ($= \frac{194,232}{758,636}$) to 32% ($= \frac{389,136}{1,231,852}$).

[1]The analysis is conservative in that obscure Website aliases are not identified but is liberal in that empty search-parts are not excluded.

Web information location service usage can be measured by the overall proportion of search-query-clicks or the *average search-query proportion* which is defined as,

$$\text{average search-query proportion} = \frac{\sum \text{search-query-click rate}}{\sum \text{click rate}}, \text{ search-query-clicks per click.}$$

Compared with query-clicks, the average search-query proportion increased only slightly from 11% ($= \frac{86,853}{758,636}$) to 12% ($= \frac{151,761}{1,231,852}$).

The overall proportion of search-sessions or the *average search-session proportion* is defined as,

$$\text{average search-session proportion} = \frac{\sum \text{search-session rate}}{\sum \text{session rate}}, \text{ search-sessions per session.}$$

The average search-session proportion remains the same at 49% ($= \frac{10,483}{21,366}$ and $\frac{12,311}{25,192}$) during each of the study-years even though the average proportion of query-sessions (or $\frac{\sum \text{query-session rate}}{\sum \text{session rate}}$, query-sessions per sessions) which has also already been seen, increased from 68% ($= \frac{14,551}{21,366}$) to 75% ($= \frac{18,996}{25,192}$).

The disparity between the increase in the proportion of query-sessions and the lack of an increase in the proportion of search-sessions may be explained by a structural change in the Web whereby search-parts, and hence query-sessions, are used more prevalently by Websites independently of any potential change in the use by student-users of Web information location services. (The study-year frequencies relating to clicks, query-clicks, search-query-clicks, sessions, query-sessions and search-sessions proportion are given in Tables C.1 and C.2 in Appendix C.)

The popularity of Web information location services is measured by counting the number of different individual search-users who submit search-queries to each service during any given period. Table 5.1 shows the ranks of the fourteen principal Web information location services or the most popular services according to their popularity during both study-years.

Service	**Popularity** (search-users during period)			**Search-query**
	study-year one (1998-1999)	study-year two (1999-2000)	both study-years	both study-years
Yahoo!	737	911	983	59,233
Google	541	811	911	32,045
AltaVista	499	744	851	82,217
Excite	588	124	622	12,894
Infoseek	373	427	596	18,349
Lycos	422	189	521	9,720
Netscape	399	202	507	3,460
LookSmart	225	113	302	2,414
AskJeeves	21	242	248	6,124
SavvySearch	52	206	238	4,415
InFind	73	169	221	2,612
Go	116	126	215	2,636
HotBot	63	60	110	1,206
WebCrawler	56	47	97	1,289
Totals:	1,017 student-users	1,035 student-users	1,050 student-users	238,614 clicks

Table 5.1: Popularity of Web information location services

A collective-popularity analysis (see Chapter four) shows that during both study-years all 1,050 student-users used[2] at least one of Yahoo!, Google, AltaVista, Netscape or Excite.

The popularity ranking confirms the dominance of the leading brand names both in respect of search-query-click frequency and their growth in popularity from study-year one to study-year two. For example the leading three services have captured over 70% of all search-query-clicks. In contrast, the popularity of the lesser brands such as Excite and Lycos has declined. (Precise comparisons are not reliable since a service may not have been available during the whole two year period[3]).

It appears that search-users change the Web information location service which they use. This interpretation of the Web log is supported by the increase in relative-collective-popularity. During study-year one all the 1,017 search-users used at least one of the top *twelve* ranked services while during study-year two all of the 1,035 search-users who used Web information location services used at least one of the top *nine* ranked services. 757 and 963 search-users used either or both of Yahoo! and Google during study-years one and two respectively. (During both study-years all 1,050 student-users used at least one of the top seven ranked services.)

The major Web information location services can thus be said to be very popular since during each study-year about 98% of student-users use them. However in half of all sessions, they are not used at all. The conjecture that every student-user makes *some* use of a Web information location service during the two study-years is proven but overall, by-session, Web information location services are no more likely to be used than not.

Are some student-users more likely than others to use Web information location services? This possibility is now investigated and it is found that, for example, conformant student-users are more likely (session for session) to use these services.

The student-user's *search-session rate* is the count of the number of search-sessions by the student-user during a study-year. The mean search-session rate during study-year two increased ($p < .01$, $z = 3.01$) from 10.0(0.4) search-sessions per study-year during study-year one to 11.7(0.4) search-sessions per study-year during study-year two. However, as Figure 5.1 illustrates, there is a great variety in student-users' search-session rates and the frequency distributions are (unsurprisingly) asymmetric so that they resemble the distributions of session rate discussed in Chapters three and four. It is also inevitable, because of the constraint on the frequency distribution imposed by the underlying session rate distribution, that the two study-years'

[2] Inspection of the Websites involved reveals that Google was mainly accessed through Yahoo! and that Netscape was providing re-branded access to other services (see Table 4.7 on page 138).

[3] The 'search engine' market has undergone a 'shake-out' since the study. Individual service offerings are much revised and several brands no longer exist.

search-session rate distributions resemble each other (because of the similarities of the underlying session rate distributions).

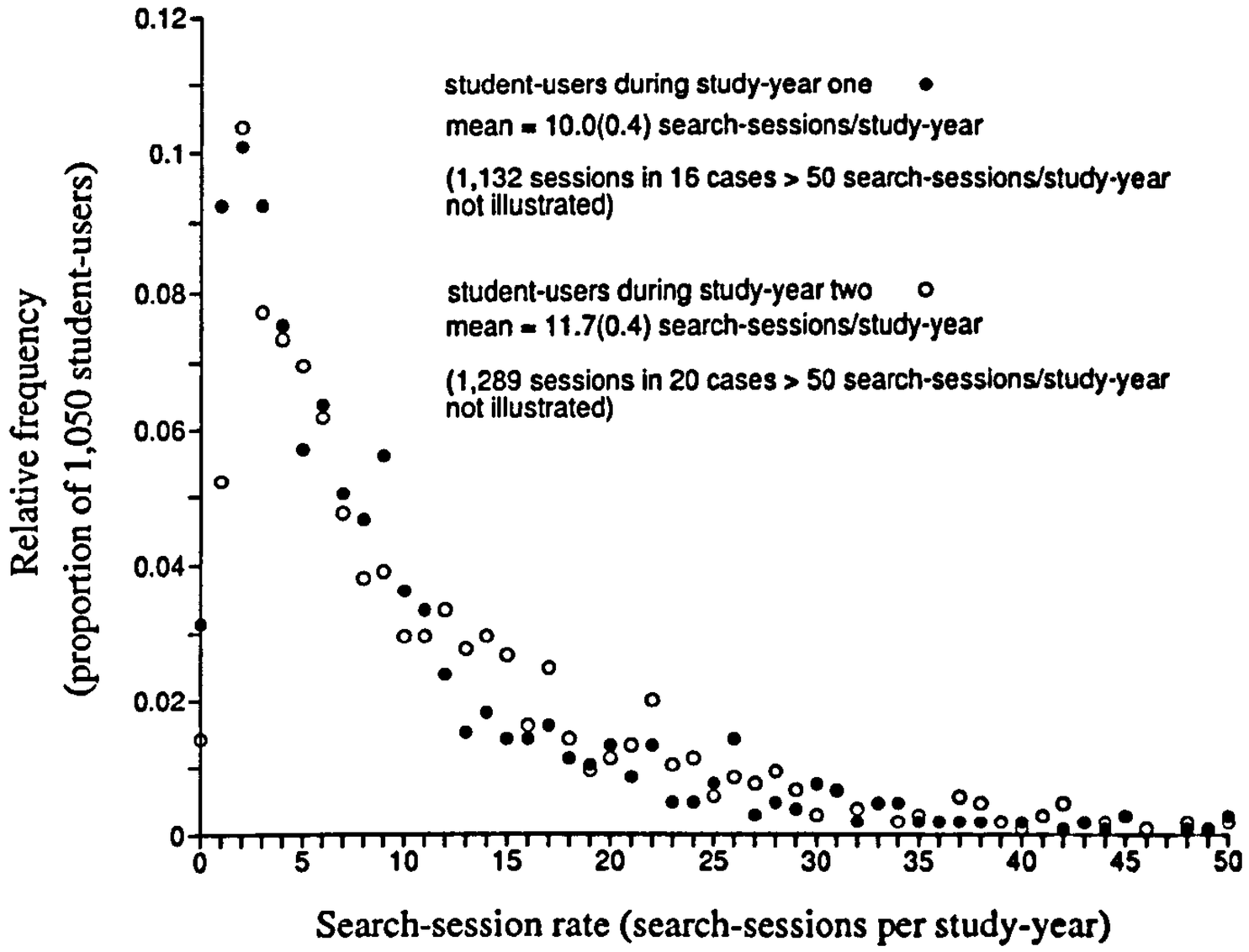

Figure 5.1: Relative frequency distributions of student-user's search-session rate

In a (banal) sense, a student-user with a search-session rate of 50 search-sessions per study-year is five times more likely to use Web information location services than a student-user with a search-session rate of ten search-sessions per study-year. It is of more interest to discover if different groups of student-user are uniform or if they differ in their *proportion* of search-sessions (that is $\frac{\sum \text{search-session rate}}{\sum \text{session rate}}$). For example, do men student-users have proportionately smaller (or larger) search-session rates than women student-users?

Table 5.2 is a cross-tabulation of sessions during study-year two[4] organised to investigate this question. Sessions are classified by the gender, session-rate and conformance user-attribute of the student-user whose session it is and by whether or not each session is a search-session. Session frequencies are entered in each table cell so that, for example, (during study-year two) there are 7,858 sessions which are search-sessions from men student-users.

[4] The equivalent cross-tabulation in respect of study-year one is given in Table D.1 in Appendix D.

User-attribute	Sessions	
	search-sessions	not search-sessions
gender		
men	7,858 sessions	7,365 sessions
women	4,453 sessions	5,516 sessions
session-rate		
smaller	3,931 sessions	3,422 sessions
larger	8,380 sessions	9,459 sessions
conformance		
conformant	2,288 sessions	1,875 sessions
eclectic	10,023 sessions	11,006 sessions
Study-year two	12,311 sessions	12,881 sessions

Table 5.2: Cross-tabulation of session frequency by user-attribute partition and 'searching' during study-year two

The χ^2 test of independence tests the association between each classification attribute and whether or not sessions are search-sessions. The test statistic in respect of each user-attribute is given in Table 5.3. In each case χ^2 exceeds the critical value ($\chi^2_{1;0.01} = 5.02$). Hence the null hypothesis that whether or not sessions are search-sessions is independent of, for example gender, is rejected. It is concluded that the likelihood of a session being a search-session depends upon student-user gender and whether or not the student-user has a smaller/larger session-rate or is conformant/eclectic.

Study-year	User-attribute		
	gender	session-rate	conformance
study-year one	$\chi^2 = 83$	$\chi^2 = 64$	$\chi^2 = 250$
study-year two	$\chi^2 = 116$	$\chi^2 = 87$	$\chi^2 = 73$

Table 5.3: Cross-tabulation of χ^2 statistic for independence of 'searching' by study-year and user-attribute

Inspection of Table 5.2 on the facing page shows that, for study-year two, search-sessions from men, and from smaller session-rate and conformant student-users are over represented. The same is found in respect of study-year one using the frequencies given in Table D.1 in Appendix D. Hence the associations are *consistent* or repeated during each of the study-years.

Thus, although overall about half of all sessions involve the use of Web information location services, the sessions from men student-users, from student-users with smaller session-rates and from conformant student-users are all more likely to be search-sessions. Hence student-users with these attributes are more likely (session for session) to make use of Web information location services.

The focus of the investigation into how student-users use Web information location services turns next from search-sessions to search-queries. A sample of search-queries (the AltaVista-Excite sample) is obtained from the search-parts associated with search-query-clicks submitted to either the AltaVista or Excite services. These are now analysed.

5.3 Web search-query analyses

The investigations reported in this Section examine the AltaVista-Excite sample. There are four subsections which consider firstly the composition of the sample and then the occurrence of search-users' search-queries. The penultimate subsection is an analysis of search-queries and search-term frequencies and the concluding subsection compares the AltaVista-Excite sample with the published surveys of Excite search-queries.

During both study-years there are 238,614 search-query-clicks or Web requests to one of the fourteen principal Web information location services. 82,217 and 12,894 (together = 40%) of these search-query-clicks are to either the AltaVista or Excite services and form the basis of the AltaVista-Excite sample.[5]

The AltaVista-Excite sample

The Web log captures the client-side component of the client/server dialogue. On each occasion that a search-user submits a search-query the conditioned url-strings encode, using a protocol particular to the service concerned, the composition of the search-query, the service response and navigational information for example moving to another results page. The composition of the search-query includes the encoded *search-terms* or 'words' which have been submitted by the student-user.

The AltaVista and Excite services were selected from among the principal Web information location services because of their popularity and because the search-part encoding protocol which is used is straightforward to decode. Server-side query investigations of both these services are also included in the literature (see Chapter two) so that results can be compared.

During both study-years 954 search-users submitted search-query-clicks during 9,130 search-sessions to either the AltaVista or Excite services. The total of 926 student-users and 8,543 search-sessions included in the sample is less than the totals just noted which are derived from the popularity analysis described in the previous Section. The sample numbers are smaller because they are based on decoding each search-part in the search-query-click url-strings. Only if a non-null search-query is present is the search-query included. The popularity analysis in Chapter four included all clicks.

Decoding and analysing the search-parts allows the number of distinct search-queries within each search-session and distinct search-terms (or 'words') within each each distinct search-query to be counted. For example, the four decoded submissions 'information', 'information-retrieval', 'information retrieval' and 'information information' within a search-session are three search-queries which have one, one and two search-terms respectively. (The fourth submission has only one distinct search-term and is not a distinct search-query.)

Not every student-user used either AltaVista or Excite. Although the total number of search-users in the sample remains about the same (about 755 of 1,050) during

[5] Google search-queries are not included in the sample because initial analysis of the Web transaction analysis showed only a small volume of these; it was only later that the large volume of Google search-queries submitted through Yahoo! was discovered. Google's popularity is shown in Table 5.1 on page 153.

each study-year, only 584 out of the 756 search-users during study-year are present during study-year two. This reduction appears to be connected with the decline in popularity of the Excite service. Change in the popularity of different Web information location services is discussed in Section 5.2 above. However the search-queries from the 584 search-users do form a longitudinal-developmental sample since there is a sample of search-queries from each search-user during each study-year. This is investigated later in Chapter six. 124 ($= 1,050 - 926$) student-users do not appear in the sample. Possible bias among the sample search-users which might explain this is examined below.

Table 5.4 shows the 124 student-users who do not appear in the AltaVista-Excite sample classified by their gender attribute. The χ^2 test for goodness of fit is used in order to test for a gender bias, for example, are these student-users disproportionately men? The test statistic $\chi^2 = 2.21$ does not exceed the critical value of $\chi^2_{1;0.05} = 3.84$, so it is accepted ($p < .05$) that there is no gender bias among the 124 student-users not in the AltaVista-Excite sample. It therefore follows that there is no gender bias among the 926 student-users who do appear in the AltaVista-Excite sample.

	Gender user-attribute	
	men	**women**
not in the AltaVista-Excite sample	55 student-users	69 student-users
all student-users	540 student-users	510 student-users

Table 5.4: Cross-tabulation of student-user frequency for the AltaVista-Excite sample complement by gender

Similar tests[6] for bias in respect of the session-rate and conformance attributes produces the χ^2 test statistics given in Table 5.5 which shows that in each case χ^2 exceeds the critical value of $\chi^2_{1;0.05} = 3.84$. Hence non-appearing student-users tend to have smaller session-rates and tend to be conformant. Thus student-users who have a smaller session-rate or who are conformant appear during each study-year to be comparatively averse to using either the AltaVista or Excite Web information location services. This particular finding is in contrast to the general finding that smaller and conformant student-users are more likely to use Web information location services. It therefore appears that these 124 student-users are loyal users of some *other* (not AltaVista or Excite) Web information location services.

[6] The associated cross-tabulation of student-user frequencies is given in Table D.3 in Appendix D.

Study-year	User-attribute		
	gender	session-rate	conformance
study-year one	$\chi^2 = 2.2$	$\chi^2 = 15$	$\chi^2 = 30$
study-year two	$\chi^2 = 2.2$	$\chi^2 = 8.6$	$\chi^2 = 4.6$

Table 5.5: Cross-tabulation of χ^2 statistic for goodness of fit of student-users for the AltaVista-Excite sample complement by study-year and user-attribute

In the next subsection the AltaVista-Excite sample is examined by-user in order to reveal similarities and differences within these overall results in how student-users use the AltaVista and Excite Web information location services.

Search-user's search-queries

It is seen above that some student-users have a greater propensity to use Web information location services than others in the sense of the likelihood that one of their Web information seeking sessions is a search-session. The investigation now addresses the two questions of how alike search-users are in respect of the number of search-queries which they submit during a search-session, and how alike search-users are in respect of the number of search-terms which they use to compose each search-query.

The *search-query count* (search-queries per search-session) counts the number of distinct search-queries within a search-session and the *search-term count* (search-terms per search-query) counts the number of distinct search-terms within a search-query. Both of these counts measure Web information locating *usage* but they do not necessarily inform an understanding of how individual users use Web information locating services. Comparable usage data is reported in the literature and analyses of the AltaVista-Excite sample here is compared with the published Excite search-query analyses later in this Section.

A feature of this investigation is that the (anonymous) student-user is known. Therefore search-sessions can be grouped *by-user*. This allows for each search-user computation of that search-user's two characterization metrics *average search-query count* which is the ratio of the total number of search-queries to the total number of search-sessions, and the *average search-term count* which is the ratio of the total number

of search-terms to the total number of search-queries. Hence these characterizations are defined as,

$$\text{average search-query count} = \frac{\sum \text{search-queries}}{\sum \text{search-session rate}}, \text{ (search-queries per search-session)}$$

and,

$$\text{average search-term count} = \frac{\sum \text{search-terms}}{\sum \text{search-queries}}, \text{ (search-terms per search-query).}$$

Overall, as shown in Table 5.6, Web information location service usage during each of the study-years increases from 3.14 ($= \frac{13,200}{4,202}$) search-queries per search-session during study-year one to 3.48 ($= \frac{15,111}{4,341}$) search-queries per search-session during study-year two while the average search-term count (average search-term count $= \frac{\sum \text{search-terms}}{\sum \text{search-queries}}$ search-terms per search-query) remains about the same at 2.63 ($= \frac{34,718}{13,200}$) search-terms per search-query and 2.62 ($= \frac{39,544}{15,111}$) search-terms per search-query.

Study-year	Search-terms	Search-queries	Search-sessions	Search-users
study-year one	34,718 search-terms	13,200 search-queries	4,202 search-sessions	756 search-users
study-year two	39,544 search-terms	15,111 search-queries	4,341 search-sessions	754 search-users
Overall	74,262 search-terms	28,311 search-queries	8,543 search-sessions	926 search-users

Table 5.6: Cross-tabulation of search-term, search-query, search-session and search-user frequencies by study-year for the AltaVista-Excite sample

Figure 5.2 illustrates the relative frequency distributions for the search-user's average search-query count (rounded to the nearest quarter). The average search-query count for only search-users who had more than five search-sessions during the study-year is included in the illustration because of the distortion caused by small denominators. The graph shows a modal value of about three search-queries per search-session and a similarity in the asymmetric distribution for each of the study-years. The mean value for the average search-query count (over all the student-users) is greater ($2.8(0.07) > 2.6(0.06)$ $p < .05$, $z = 2.17$) during study-year two than during study-year one. Hence, as a group, search-users during study-year two submitted slightly more search-queries on each occasion when they used AltaVista or Excite compared

to during study-year one. However caution is needed when interpreting this metric since the duration of sessions is not known.

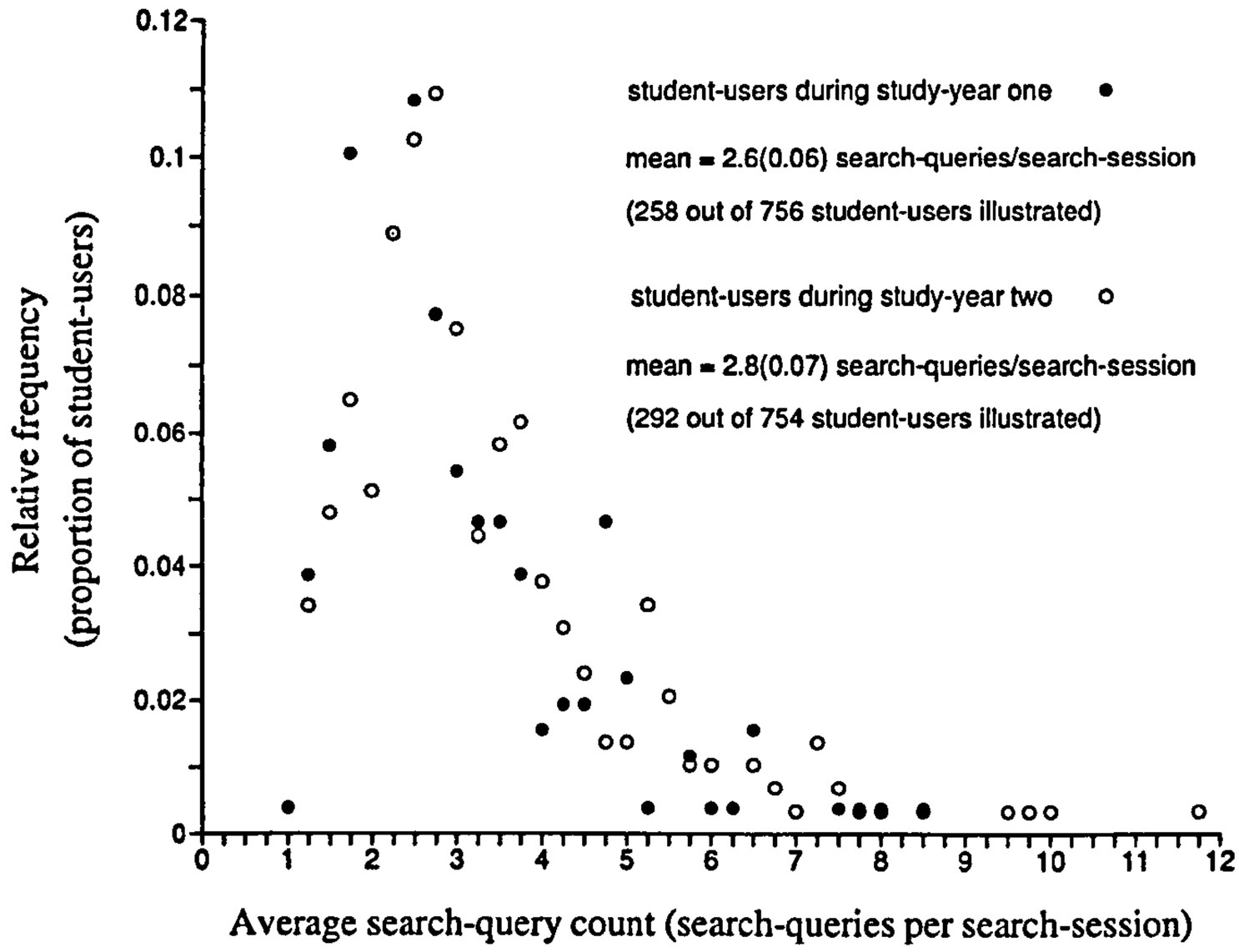

Figure 5.2: Relative frequency distributions of student-user's average search-query count

Figure 5.2 also shows that individual search-user's average search-query counts vary over a substantial range (from one up to at least eleven search-queries per search-session). However investigating this variation per se is outside the scope of this dissertation since, for example, a search-user's search-query count may be affected by search-session duration. Search-users who have longer average search-session durations may in consequence have larger average search-query counts. (The longitudinal-developmental investigation discussed in Chapter six suggests that men student-users do not change their search-query count.)

Are some search-users more likely than others to submit several search-queries during a search-session?

Search-users can be classified by both their user-attribute (gender, session-rate or conformance) and by whether or not their average search-query count is less than three. The criterion of three is arbitrary but is selected because it allows the collection of search-queries to be partitioned into those less than (or greater than) the mean average search-query count during each of the two study-years. The frequency cross-tabulations showing this are given in Tables D.4 and D.5 in Appendix D.

Bias in the classification is tested using the χ^2 test of independence and the test statistics for each user attribute are reported in Table 5.7. This shows that in each

case χ^2 is less than the critical value $\chi^2_{1;0.05} = 3.84$. Hence whether or not search-users use three or more search-queries is consistently (or during each study-year) independent ($p < .05$) of the search-user's gender, session-rate or conformance attribute.

	User-attribute		
Study-year	gender	session-rate	conformance
study-year one	$\chi^2 = 0.69$	$\chi^2 = 0.83$	$\chi^2 = 2.10$
study-year two	$\chi^2 = 1.53$	$\chi^2 = 0.00$	$\chi^2 = 0.11$

Table 5.7: Cross-tabulation of χ^2 statistic for independence of average search-query count size by study-year and user-attribute

Are average search-query and average search-term counts associated? That is, for example, do search-users who submit only a few search-queries also use only a few search-terms in these search-queries?

The mean value of search-users' average term-count remains at 2.5(0.04) search-terms per search-query during each of the two study-years. However 64 and 50 search-users during study-years one and two respectively have an *average* search-term count of just one search-term per search-query. These users are discussed later, see page 171, since the average of one search-term per search-query implies that they only ever use singleton search-queries.

Figure 5.3 illustrates the relative frequency distributions of search-users' average search-term count or the average number of distinct search-terms submitted during each distinct search-query (which like the average query-count is rounded to the nearest quarter). The singleton always search-users cause the distribution to be bi-modal but the similarity between the study-years is still evident.

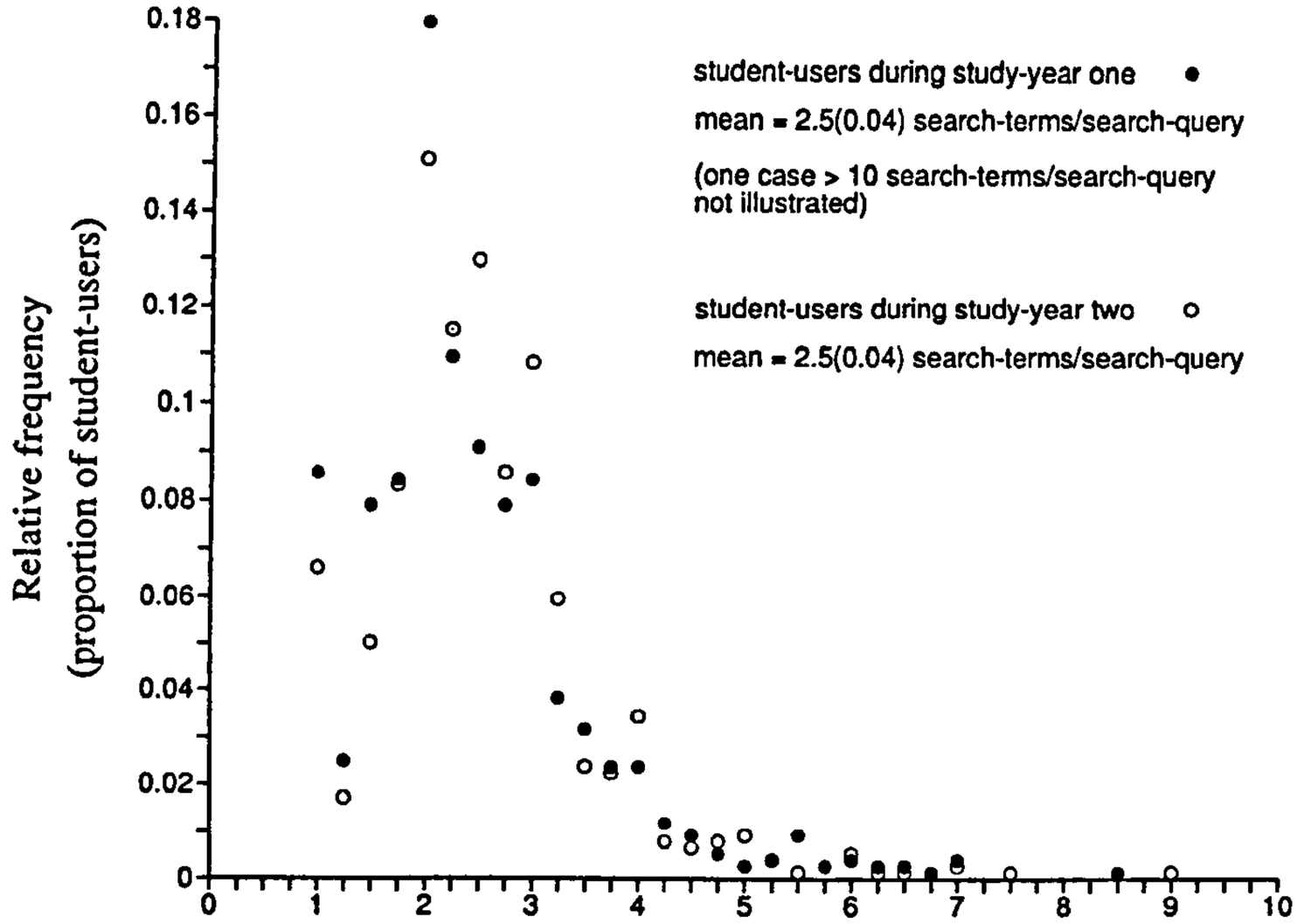

Figure 5.3: Relative frequency distributions of student-users' average search-term count

Overall therefore, student-users neither increase nor reduce the average number of search-terms which they use in a search-query. Whether or not *individual* student-users have the same average search-term count during study-year two as they do during study-year one is one of the subjects of the longitudinal-developmental investigation in Chapter six. This shows that, overall, the AltaVista-Excite search-users reduce their average search-term counts.

The scattergram in Figure 5.4 relates the average search-query count for 292 search-users during study-year two[7] to their average search-term count. These (the 292 search-users during study-year two and 258 search-users during study-year one) are the student-users who have more than five AltaVista-Excite search-sessions during a study-year. The Pearson correlation coefficients for the scattergrams are r = 0.22 and r = 0.07 for study-years one and two respectively; r is significant ($p < .05$) during study-year one but not during study-year two. Hence during study-year one there is an association between average search-query count and average search-term count but not during study-year two. The linear association between these student-user's average search-query count and average search-term count is thus not consistent during each of the study-years.

[7] The equivalent scattergram in respect of 258 search-users during study-year one is shown in Figure D.1 in Appendix D.

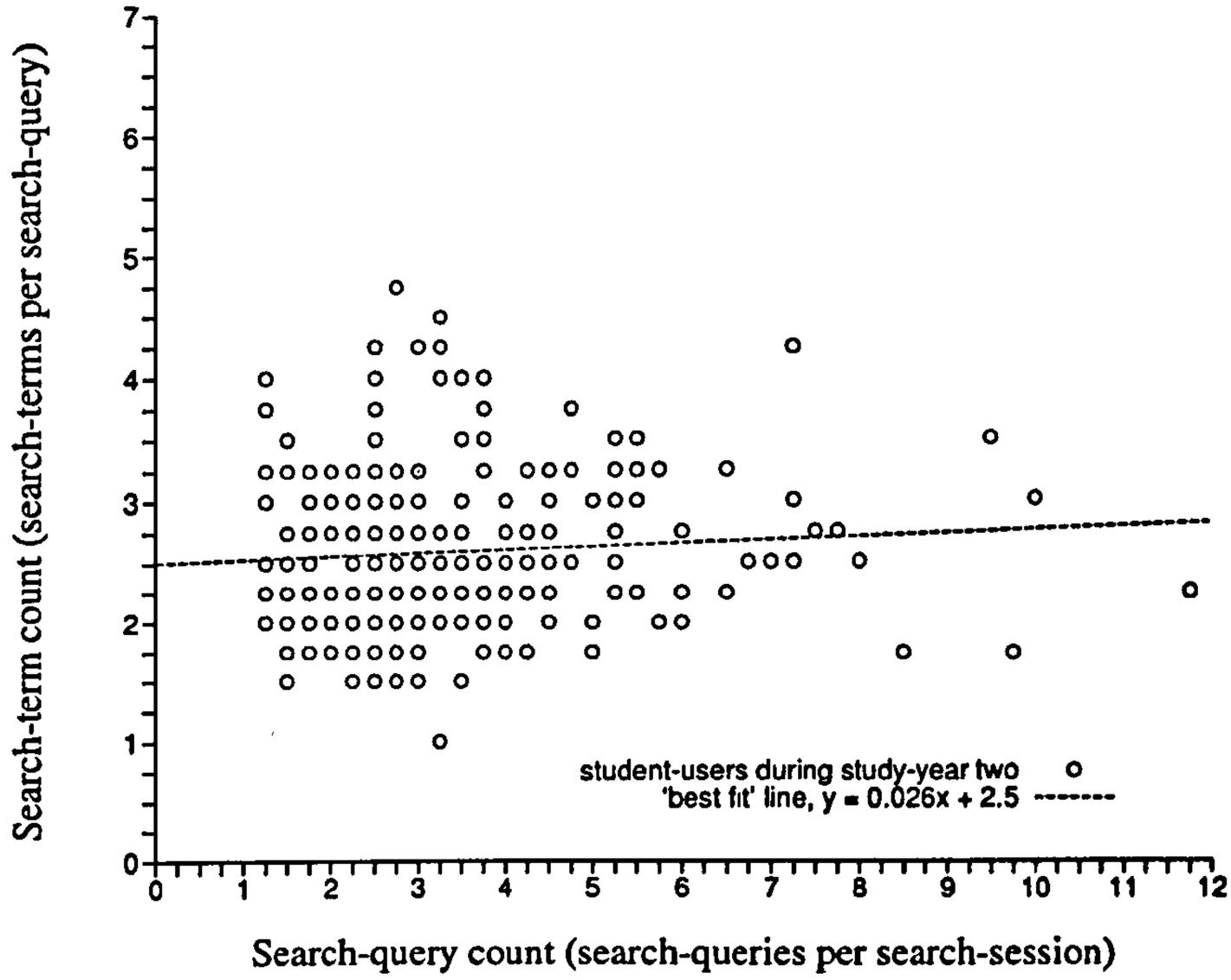

Figure 5.4: Scattergram of 292 search-user's average search-query count and search-term count during study-year two

A more inclusive (that is not excluding search-users with fewer than six search-sessions) nonparametric analysis of search-user's average search-query count and average search-term count makes use of the search-query count classification already introduced, that is search-users with average < 3 and $\geqslant 3$ search-queries per search-session. Search-users can be classified similarly by whether or not their average search-term counts are less than three.

Table 5.8 gives the frequency distribution for all the search-users in the AltaVista-Excite sample during study-year two[8] classified by the search-query and search-term criteria. The χ^2 test for independence examines the association between the classification criteria. $\chi^2_{\text{study-year two}} = 0.03$ which is less than the critical value ($\chi^2_{1:0.05} = 3.84$) so that during study-year two student-user's average search-query count and average search-term count do not depend on each other ($p < .05$). But $\chi^2_{\text{study-year one}} = 21$, so that during study-year one student-user's average search-query count and average search-term count are not independent ($p < .05$).

[8] Table D.6 in Appendix D gives the corresponding search-user frequency distribution during study-year one.

Average search-query count search-queries per search-session	**Average search-term count** search-terms per search-query	
	< 3	$\geqslant 3$
< 3	345 search-users	125 search-users
$\geqslant 3$	206 search-users	78 search-users

Table 5.8: Cross-tabulation of AltaVista-Excite search-user frequency by average search-query count and average search-term count during study-year two

The conclusion from the χ^2 test used here, which includes all the search-users in the AltaVista-Excite sample, is that average query-count and term-count are associated during study-year one but not during study-year two and hence the association is not consistent (during each study-year). This conclusion is the same as that of the linear correlation analysis (which does not include all the AltaVista-Excite sample search-users) although during study-year one search-users who have smaller search-query counts are more likely to use fewer than three search-terms.

Are student-user's average search-term counts associated with the other user-attributes which are being considered that is, gender, session-rate or conformance? Search-user classification frequencies for these attributes are set out in Tables D.7 and D.8 in Appendix D. When these frequencies are tested for independence then in each case, as shown in Table 5.9, the χ^2 test statistic is consistently less than the critical value ($\chi^2_{1;0.05} = 3.84$). Thus it is accepted that these classification attributes are independent and that there is no association between a student-user's gender, session-rate, or conformance attribute, and whether or not the student-user's average search-term count is less than three search-terms per search-query.

Study-year	User-attribute		
	gender	session-rate	conformance
study-year one	$\chi^2 = 3.09$	$\chi^2 = 0.45$	$\chi^2 = 0.13$
study-year two	$\chi^2 = 2.76$	$\chi^2 = 0.79$	$\chi^2 = 0.00$

Table 5.9: Cross-tabulation of χ^2 statistic for independence of average search-term count size by study-year and user-attribute

Hence it is found that search-user's average search-query and search-term counts range widely and overall are each independent not only of the student-user's gender, session-rate or conformance attributes but also of each other.

AltaVista and Excite search-queries and search-terms

In the previous subsection the analysis focussed on each search-user. It is seen that there is a broad range in how search-users locate information using Web information location services. It is also seen that search-user's average search-term counts are generally independent of the search-user's gender, session-rate or conformance attributes and their average search-query counts. The investigation in this subsection initially gives priority to each search-session and each search-query so that it gives them equal weight and hence mimics server-side Web information location services analyses.

The mean search-query count of 3.48(0.06) search-queries per search-session during study-year two is greater ($p < .01$, $z = 4.01$) than the mean of 3.14(0.06) search-queries per search-session during study-year one. However as mentioned previously, because investigating the search-query metric per se entails the session duration its separate study is intractable. The mean search-term count remains the same at about 2.6 search-terms per search-query during each study-year.

Figures 5.5 and 5.6 illustrate for each study-year the relative frequency distributions of the search-query and search-term counts. Both graphs show a preponderance of smaller values; most search-sessions involve only one or two search-queries (61% and 56%, study-years one and two respectively) and most search-queries (57% during each of study-years one and two) contain only one (singleton) or two search-terms.

This suggests that these two measures may be associated. A particular question is, are singleton search-queries more or less likely in smaller (< 3 search-query) search-sessions?

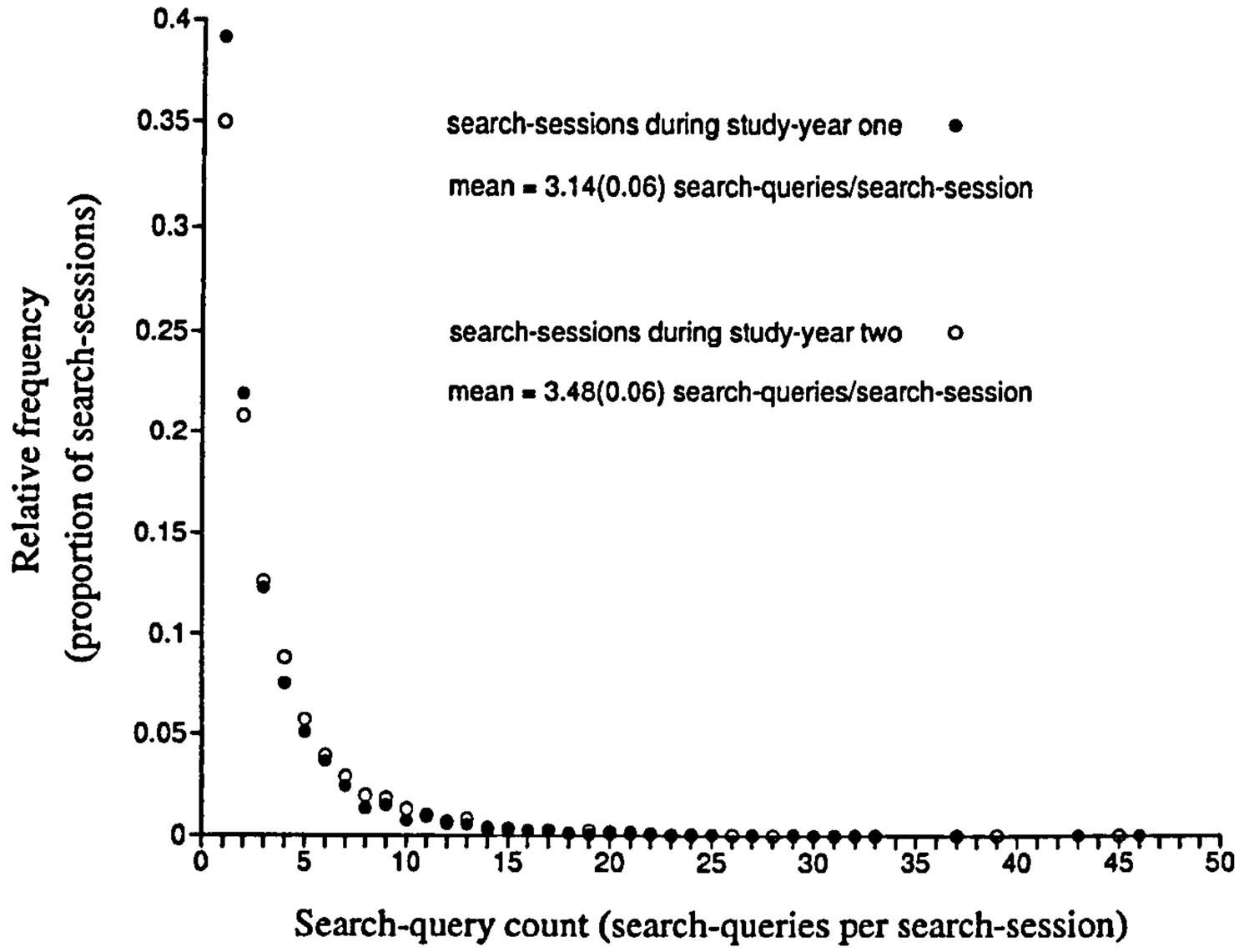

Figure 5.5: Relative frequency distributions of the number of search-queries submitted during each search-session

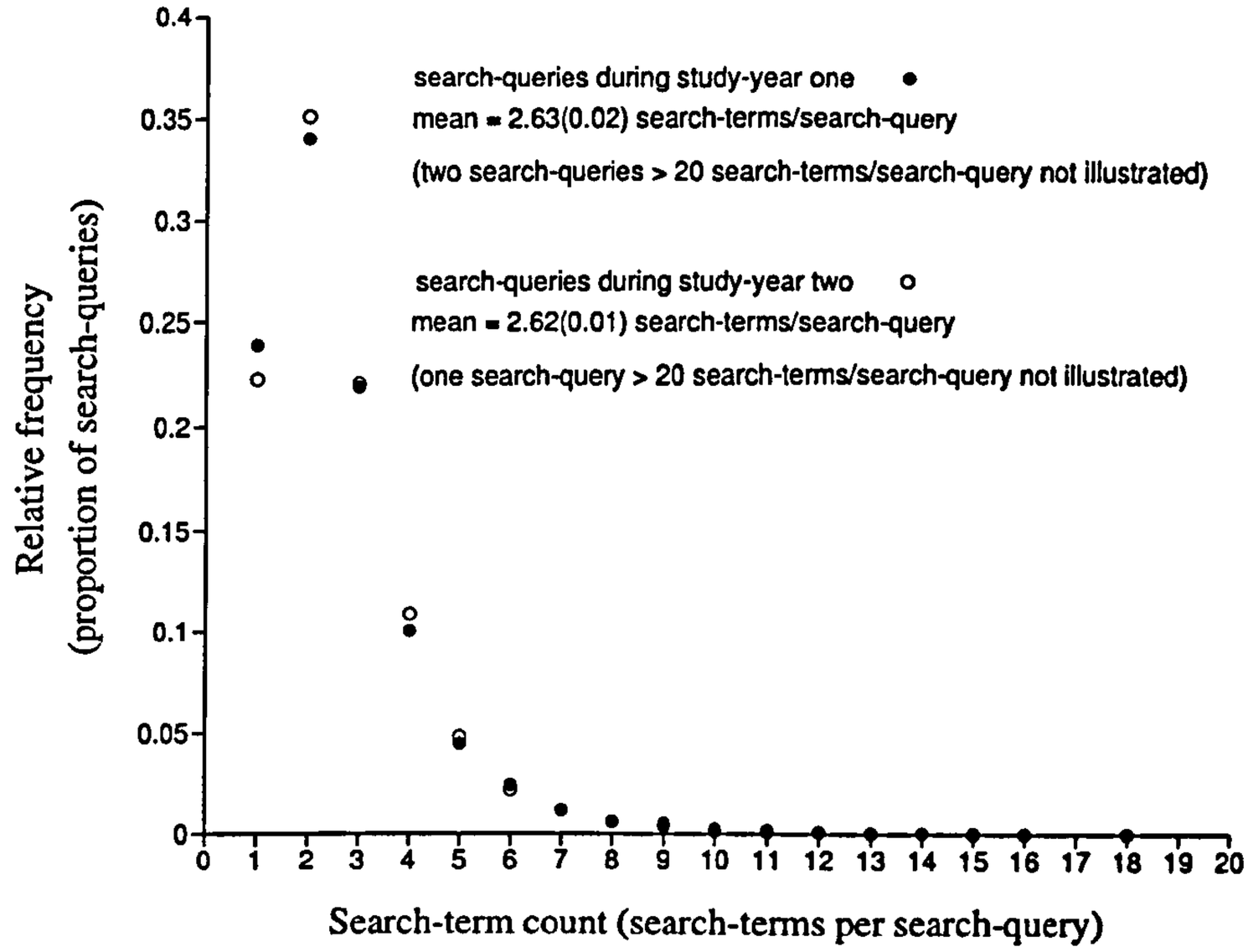

Figure 5.6: Relative frequency distributions of the number of search-terms submitted in search-queries

Table 5.10 shows 15,111 search-queries during study-year two[9] cross-tabulated according to whether or not they are singleton and whether the search-query occurs in a smaller or larger search-session. The independence of the two attributes, whether a search-query is a singleton and whether the search-session involves more than two search-queries, is tested with the χ^2 test for independence. For both study-years χ^2 exceeds ($\chi^2_{\text{study-year one}} = 106$, $\chi^2_{\text{study-year two}} = 61$) the critical value ($\chi^2_{1:0.05} = 3.84$) thus the hypothesis that the attributes are independent is rejected. Hence, whether or not singleton search-queries are submitted is associated with the size of the search-session as measured by the search-query count.

[9] The equivalent cross-tabulation in respect of study-year one is given in Table D.9 in Appendix D.

Search-session	Search-queries	
	singleton	non-singleton
smaller	908 search-queries	2,419 search-queries
larger	2,456 search-queries	9,328 search-queries

Table 5.10: Cross-tabulation of search-query frequency by search-session size and search-query count during study-year two

There is a bias towards singleton search terms in the smaller (< 3 search-queries) search-sessions during each of the study-years. Thus, consistently, smaller search-sessions are more likely to involve singleton search-queries than are larger search-sessions. (And equivalently, singleton search-queries are more likely to occur in smaller search-sessions. An aspect outside the scope of this investigation is that student-users who submit a singleton search-query must therefore have a predilection to smaller search-sessions.)

There is also a consistent bias towards singleton search-queries in search-queries from search-users with smaller session-rates revealed by the χ^2 test of independence applied to search-queries classified by search-user gender, session-rate or conformance attribute and whether singleton or not. The search-query frequencies cross-tabulated by search-user gender, session-rate and conformance attribute, and whether singleton or not is given in Tables D.10 and D.11 in Appendix D. The χ^2 test statistics are set out in Table 5.11 which, in addition to the consistent session-rate bias ($p < .05$, $\chi^2 > \chi^2_{1:0.05}$ during each study-year) shows that whether or not a search-query is singleton is consistently independent ($p < .05$, $\chi^2 < \chi^2_{1:0.05}$ during each study-year) of the student-user's gender. The relationship between the proportion of singleton search-queries and whether the student-user is conformant or eclectic is not consistent. During study-year one singleton search-queries are more likely ($p < .05$, $\chi^2 > \chi^2_{1:0.05}$) from conformant search-users but that during study-year two the use of singleton search-queries is independent ($p < .05$, $\chi^2 < \chi^2_{1:0.05}$) of whether or not the student-user is conformant.

Study-year	User-attribute		
	gender	session-rate	conformance
study-year one	$\chi^2 = 2.2$	$\chi^2 = 5.7$	$\chi^2 = 5.6$
study-year two	$\chi^2 = 0.40$	$\chi^2 = 16.9$	$\chi^2 = 0.00$

Table 5.11: Cross-tabulation of χ^2 statistic for independence of singleton search-query occurrence by study-year and user-attribute

During each study-year more than 6% of AltaVista-Excite search-users (see Figure 5.3) or 64 and 50 student-users respectively, have an *average* search-term count of only one, that is *all* of their search-queries are singleton. Are these student-users a distinct group in respect of the gender, session-rate and conformance attributes and do they distort the AltaVista-Excite sample analyses?

Table 5.12 shows, for example, search-user frequencies during study-year two classified by both gender and average search-term count. The χ^2 test of independence shows that there is no gender bias ($p < .05,\ \chi^2 = 1.75 < \chi^2_{1:0.05}$) in respect of the 50 search-users who used only singleton search-queries during study-year two. This is also the case during study-year one ($p < .05,\ \chi^2 = 2.95 < \chi^2_{1:0.05}$).

Student-user gender	Average search-term count	
	singleton	> one search-term per search-query
men	32 student-users	375 student-users
women	18 student-users	329 student-users

Table 5.12: Cross-tabulation of AltaVista-Excite search-user frequency by gender and average search-term count during study-year two

The search-user frequencies classified by the other attributes (session-rate and conformance) and average search-term count are similarly tested for independence.[10] During study-year two there is no session-rate ($p < .05,\ \chi^2 = 2.58 < \chi^2_{1:0.05} = 3.84$)

[10] The cross-tabulations are given in Tables D.12 and D.13 in Appendix D.

or conformance ($p < .05,\ \chi^2 = 0.21 < \chi^2_{1:0.05} = 3.84$) bias but during study-year one singleton search-users are biased in respect of both session-rate ($p < .05,\ \chi^2 = 14.25 > \chi^2_{1:0.05} = 3.84$) and conformance ($p < .05,\ \chi^2 = 4.77 > \chi^2_{1:0.05} = 3.84$).

During study-year one search-users who have smaller session-rate or are conformant are both over represented among those search-users whose AltaVista-Excite search-queries are always singleton.

However since any bias is not consistent then it cannot be said that these search-users form a distinct group.

This Section concludes with a direct comparison of the published Excite server-side analyses of the general public's search-queries and the AltaVista-Excite sample analyses undertaken here.

The AltaVista-Excite sample compared with Excite search-query surveys

Server-side analyses of samples of search-queries from the general public received by the Excite Web information location service (the Excite samples) are discussed in Chapter two. The number of search-terms in each search-query (search-term count) in the AltaVista-Excite sample of this investigation can be reliably compared with that of the Excite samples. However the number of search-queries submitted during a search-session cannot be reliably investigated in the absence of a reliable definition of the duration of a session. For example, each AltaVista-Excite search-session which is used here encompasses the full day which is available to the search-user while the Excite session appears to relate only to a period of continuous Web information seeking activity involving the Excite Web information location service. It is therefore not surprising that the search-user's mean search-query count here exceeds an individual Excite search-session (since each AltaVista-Excite search-session could be several Excite search-sessions).

Like the Excite surveys, the mean search-term count in the AltaVista-Excite sample remains stable over time. But the Excite cumulative distributions show an *increase* (from 26% to 30%) in singleton search-queries while search-users in the AltaVista-Excite sample *reduce* (from 23.9% to 22.3%) the proportion of their singleton search-queries.

Table 5.13 compares both search-query counts and search-term counts. Both mean values and cumulative relative frequency distributions are given. The mean search-term count of the AltaVista-Excite sample here, 2.6 search-terms per search-query during each study-year, is greater than that found in the Excite samples which

are each 2.4 search-terms per search-query. The proportion of singleton search-queries in the Excite samples (26% and 30%) are correspondingly greater than in the AltaVista-Excite sample (24% and 22% during study-year one and two respectively). An interpretation of this is that Excite users from the general public are more prone to use singleton search-queries than are the student-users who are included in the AltaVista-Excite sample.

	Excite studies		AltaVista-Excite sample	
	16 Sept. 1997	20 Dec. 1999	study-year one (1998-1999)	study-year two (1999-2000)
Sample size search-queries	>1 million	>1 million	13,200	15,111
Search-query count (search-queries/search-session) mean value	2.5	1.9	3.14(0.05)	3.48(0.06)
cumulative distribution				
1	48%	60%	39.1%	35.0%
2	69%	80%	61.0%	55.8%
3	100% (3+)	100% (3+)	73.3%	68.4%
4+			100%	100%
Search-term count (search-terms/search-query) mean value	2.4	2.4	2.63(0.02)	2.62(0.01)
cumulative distribution				
1	26%	30%	23.9%	22.3%
2	58%	67%	57.9%	57.4%
3	100% (3+)	100% (3+)	79.8%	79.4%
4+			100%	100%

Table 5.13: Comparison of the AltaVista-Excite sample with public Excite samples

However the mean search-term count of 2.6 search-terms per search-query (and the Excite sample's 2.4 search-terms per query) is a by-session metric not a by-user metric and is larger than the 2.5 search-terms per search-query by-user metric noted above. The excess of 2.6 over 2.5 search-terms per search-query is explained by the proportion of student-users who submit a disproportionately large number of search-queries each with a *large* (> 2) number of search-terms. If a disproportionately large number of *small* (< 3) search-queries were submitted instead then the relationship between the by-session and by-user metrics would reverse.

Therefore, it is known that each student-user in the AltaVista-Excite sample typically uses 2.5 search-terms, but it is not known whether each user in the Excite sample typically uses > 2.4 or < 2.4 search-terms, that is whether or not there is an excess or deficit of small (< 2 search-terms) search-queries in the Excite samples. If it is the former then the difference between the two user samples (AltaVista-Excite and Excite) are reduced, if it is the latter then the difference is magnified.

5.4 Summary and discussion

The investigation described in this Chapter examines how student-users use Web information location services. These services are defined as being the fourteen principal Webhost services to which student-users submit 'information retrieval' type search-queries. When a student-user is doing this, he (or she) is 'searching'.

At some time during the two years of the study every student-user used at least one of the top seven most popular Web information location services. However despite this universal popularity overall, 'searching' occurs in only half of all Web information seeking sessions. Men student-users, student-users who have smaller session rates and conformant student-users all 'search' in more than half of their sessions. On the other hand, women student-users, student-users who have larger session rates and eclectic student-users 'search' in less than half of their Web information seeking sessions. 'Searching' does not increase over time so that any increase in user's systematicity is not evident as an increase in the use of IR type queries.

The greater 'searching' by the smaller and by the conformant student-users can be interpreted as them carrying out more territory (or Website vocabulary) building. They are thus developing their personal Web information infrastructure. In contrast the larger student-users and the eclectic student-users are interpreted as being already more developed in their territory building so that their niche information needs are more likely to be satisfied from within their existing territory. It was seen previously that the apparent gender difference may be explained by the gender/session rate association among student-users.

Search-queries from 926 (out of 1,050) student-users (the AltaVista-Excite sample) are analysed. These search-queries are all those which are submitted to the AltaVista or Excite Web information location services during both study-years. The particular benefit of analysing these search-queries is that it allows the population of 1,050 student-users in this investigation to be compared with other populations of Web users.

On average, student-users submit 2.5(0.4) search-terms (the search-term count) in each of their search-queries. This value remains the same during each study-year and thus supports the assumption that typical session durations and information tasks are the same during each study-year. There is a preponderance of smaller search-queries which is seen in both the by-user and by-session analyses. However search-term counts and search-query counts (the number of search-queries during a session) vary widely. These counts are independent not only of the user-attributes but also of each other.

Over 73% of student-users have an average search-term count of less than three search-terms per query and 57% of search-queries contain less than three search-terms. The use of only one or two search terms in a search-query is ubiquitous within the AltaVista-Excite sample and is unaffected by the student-user's gender, session-rate or conformance attribute. Nor is there any consistent (during each study-year) association generally between the number of search-queries submitted during a search-session and the number of search-terms in each search-query. The special case of singleton search-queries, that is search-queries with just one search term, is however associated with smaller search-sessions.

Server-side analyses of samples of search-queries from the general public received by the Excite Web information location service (the Excite studies) have been published (see Chapter two). The search-term counts in the AltaVista-Excite sample of this investigation can be reliably compared with that of the Excite studies. However the number of search-queries per se submitted during a search-session cannot be reliably investigated in the absence of a reliable definition of the duration of a session. The session definition used in the Excite studies implies a shorter session than the daily definition used here.

The mean search-term count here, 2.6 search-terms per search-query during each study-year, is greater than that found in the Excite samples which are each 2.4 search-terms per search-query. The proportion of singleton search-queries in the Excite samples (26% and 30%) are correspondingly greater than in the AltaVista-Excite sample (24% and 22% during study-year one and two respectively). An interpretation of this is that Excite users from the general public are more prone to use singleton search-queries than are the student-users who are included in the AltaVista-Excite sample here.

Although the search-term count is slightly greater the similarity, in particular the stability of mean search-term count over a prolonged period seen with both user populations, suggests that the 'searching' by student-users resembles that of the general public as seen by the major Web information location services.

Since 'searching' occurs in only a minority of Web information seeking sessions and, for those student-users who make a greater use of the Web, Web information seeking by 'searching' is reducing, then 'searching' does not appear to be a good basis for describing generally how student-users locate Web information. As previously however, this conclusion is based on considering groups of users rather than individuals.

6

How do novices seek Web information?

6.1 Introduction

The analyses of Web information seeking activity reported in this Chapter investigate *change* over time, that is the analyses are longitudinal and compare Web information seeking activity between the two extended surveys of study-year one and study-year two. The focus of attention in this Chapter returns to the student-users and the seven user-characterization metrics on which the interpretation of the Web log is based. In addition there is also a longitudinal examination in respect of how Web information location services are used and which uses the additional four Web information location service characterization metrics described in Chapter five.

The changes in student-users' user-characterizations reported in Chapters four and five are for *groups* of student-users which are defined by some user-attribute, for example gender. Hence men student-users' and women student-users' Web information seeking activity can be analysed longitudinally so that the groups' activity during study-year two is compared with the groups' activity during study-year one. The student-users during study-year two in this example are the *same* student-users as during study-year one but the longitudinal comparison being made is for the *group* not for individuals. If the change by some student-users within the group compensates for that by other student-users within the group then no net change for the group as a whole will be found.

When the change over time in *individuals* is studied as in this Chapter, the analysis is referred to as *longitudinal-developmental.* Hence a pre-requisite for undertaking a longitudinal-developmental study is that individuals are monitored over time. This requirement is satisfied here since the (anonymous) identity of each student-user is known consistently throughout the two years of the study. The technique of

conditional-regression facilitates longitudinal-developmental analysis since it models the change effect as a function which relates the user-characterization at a later time to the user-characterization at an earlier time.

The rationale of the *conditional analysis* (see Chapter three) undertaken here is that, in the absence of any change, an individual student-user's user-characterization during study-year two will be the same as that during study-year one. If there is any individual change then the user-characterizations will be different. The conditional-regression used also presumes that the magnitude of any change effect is *proportional* to the magnitude of the individual's user-characterization during study-year one. That is, a longitudinal-developmental phenomenon will have a greater effect on individuals who are 'large' with respect to the user-characterization in question than on 'small' individuals.

The difficulty of interpreting the findings reported in this Chapter is discussed in Chapter one. There are two contextual phenomena in particular which are recognized. The first is *structural* change, that is a change in the structure of the Web which changes its information seeking affordances and thereby affects how student-users locate Web information. Although it may not always be possible positively to identify change which is structural, non-structural change phenomena will be evident by their differential effects, that is the change will be evident in one group of student-users but not in another. For example, the conditional analysis of session-rate for men student-users and women student-users discussed in Chapter three shows that both men and women student-users increase their session-rate. Thus, if session-rate were being considered as a candidate for being a structural change phenomenon, then the lack of a differential change effect amongst the gender groups (both increase) would raise the possibility of the session-rate increase being structural. Hence it may be that at least some of the increase should be explained as being due to a structural change in the Web rather than a developmental change in the individual.

A particular interpretation of the conditional analysis of session-rate by gender demonstrated in Chapter three relates to Web information seeking *proficiency* or *expertise.*[1] The notion of change in proficiency is examined more fully in the next Section, but for example, is it reasonable to use session-rate as an indicator of expertise? That is, should users who have a larger session-rate be taken also to have greater expertise? If this were the case, then the conditional analysis of session-rate by gender implies that women increase their expertise over time significantly more so than men. Further, if some of the increase in session-rate is structural, then this could imply that men do not increase their expertise as indicated by session-rate over

[1] What expertise *is* is beyond the scope of this work but its broad properties are generally accepted.

time while women do. Since this makes a nonsense of what is commonly understood by expertise then it is concluded that session-rate is a poor indicator of expertise.

The second contextual phenomenon affecting interpretation of these longitudinal-developmental findings is that of *novice* Web users. The term as used here is in contrast to *having experience* (and vice versa) so that neither description entails any notion of possessing or not-possessing expertise in locating Web information resources (see Chapter two). The *novice-effect* which arises from the assumption that the rate of change of proficiency of a user reduces with time is used to identify potential indicators of Web information seeking proficiency. This is now discussed.

The novice-effect

Experience, or more precisely the duration of the opportunity[2] which a student-user has had for using the Web, is an attribute which can be evaluated for each student-user. Since the cohort of each (anonymous) student-user is known then a convenient measure is the number of previous (academic) years that a student-user has been at the institution. For the student-users during study-year one this is either zero or one. During study-year one, the 1998-cohort student-users are referred to as being novices since they have no previous experience.

The title of this Chapter suggests that there is a difference in the Web information seeking activity of novices and more experienced student-users. It is firstly presumed that some users of the Web are more proficient at locating information than others (the notion of proficiency is left deliberately vague). It is then hypothesized that,

a) novice student-users become more proficient as they gain experience, and that,

b) the Web information seeking activity of more proficient users can be distinguished from that of less proficient users, that is, there are observable indicators of proficiency.

Hence it should be possible to distinguish the Web information seeking activity of novice student-users from that of experienced student-user. Moreover, as novices become experienced the differences between them and more experienced users will reduce. That is, *change* in an individual's Web information seeking activity will be most evident as a novice gains experience. This is referred to as the *novice-effect* and is equivalent to hypothesizing the existence of a 'learning-curve'.

[2] Although some student-users may have had some form of access to the Web prior to their registration with the institution, it is most unlikely that any will have enjoyed the unrestricted high-speed access which the UK academic network provides.

The novice-effect provides a rationale for identifying (or eliminating) Web information seeking user-characterizations which might indicate proficiency in locating Web information. User-characterizations where the change for the novice group is an exaggeration of the change among the experienced group are possible indicators of proficiency. If the change in the user-characterization among the novices is not exaggerated compared to the experienced student-users then the user-characterization does not indicate proficiency. (In other words the exaggeration is a necessary but not-sufficient condition for the user-characterization to indicate proficiency.)

For example, student-users overall increase their average session click rate so it might be hypothesized that as student-users become more proficient so they make more Web requests during their sessions. The null hypothesis is that novice student-users increase their session click rate no more than more experienced student-users.

The conditional-regression slopes in respect of average session click rate for the novices and experienced groups, as illustrated in Figure 6.1, are 1.66(0.01) and 1.36(0.01) respectively. Hence the increase among the novices is exaggerated ($p < .00$, $z = 21$) as required and the null hypothesis is rejected. Thus it is accepted that novices do increase their average session click rate more than the experienced users and that the average session click rate shows a novice-effect.

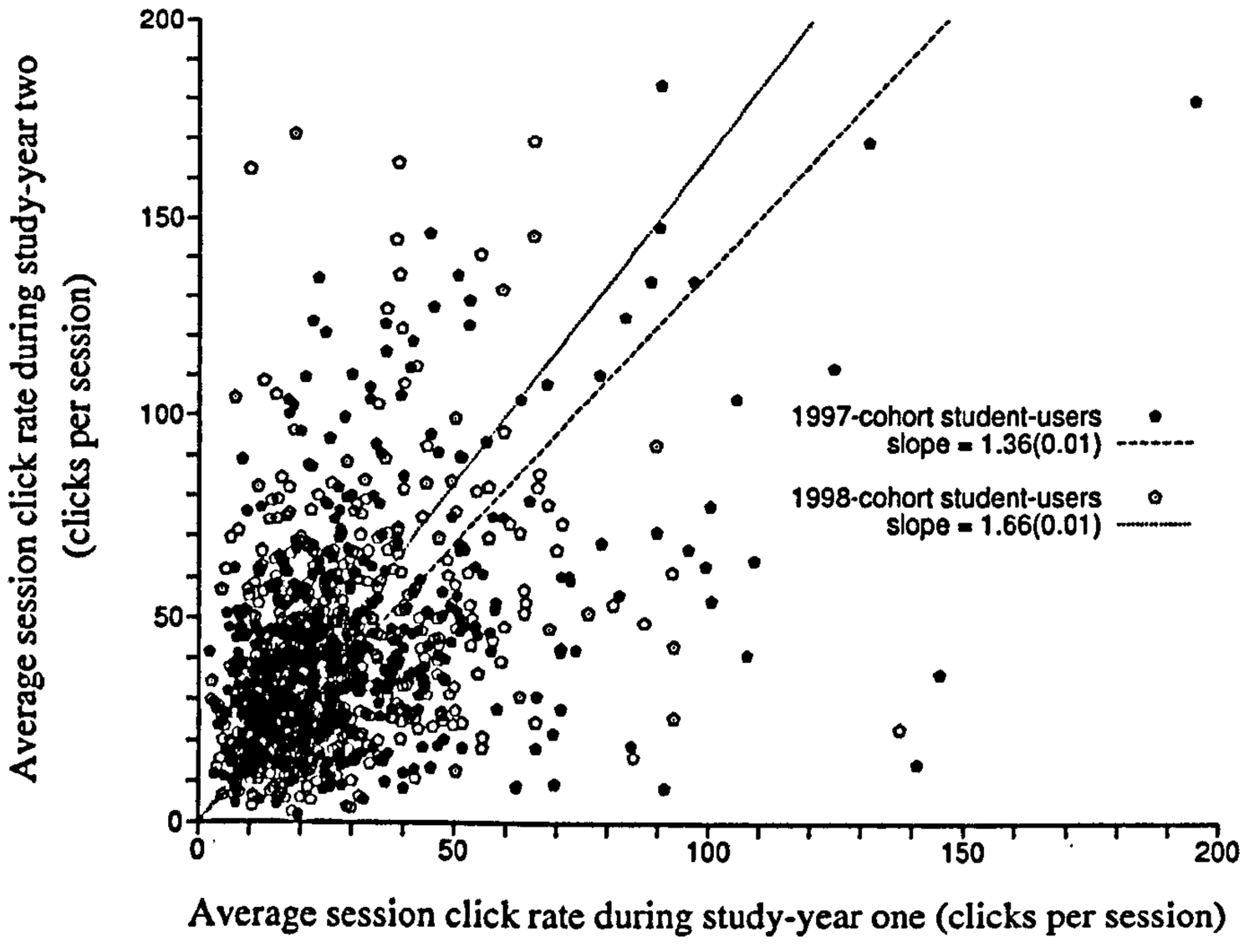

Figure 6.1: The *novice-effect* - conditional distribution of student-user's average session click rate by-cohort (range illustrated up to 200 clicks per session)

This Chapter comprises two principal Sections together with a concluding **Summary**

and discussion. The principal Sections are **How do student-users change their Web information seeking activity?** and **How do student-users change their use of Web information location services?**

The conditional analyses in **How do student-users change their Web information seeking activity?** use the seven user-characterization metrics which are introduced in Chapter four in order to characterize how student-users locate Web information. The conditional analyses also make use of user-attributes to partition the student-users into groups. Conditional analysis of the range of session-conformance produces a dispersed clustering in the conditional distribution. Because of this clustering, which is associated with the bi-modal distribution seen in Figure 4.15, a nonparametric analysis is used to support the conditional-regression technique.

In **How do student-users change their use of Web information location services?** the conditional analyses use the Web information location service user-characterizations, see Chapter five, and which describe how student-users use the principal Web information location services.

6.2 How do student-users change their Web information seeking activity?

How individual student-users change how they locate Web information is now analyzed conditionally in order to investigate individual-developmental changes in the user-characterizations. As well as comparing the conditional-regression of student-users by cohort in order to identify any novice-effect, the student-users are also compared by gender and by their *joint-session-rate* or their total session frequency during both study-years. In respect of the joint-session-rate, the student-users are partitioned by the mean[3] into 686 student-users *smaller* than average and 364 student-users *larger* than average.

The analyses relate firstly to changes in two user-characterizations which could indicate a novice-effect while those presented secondly relate to changes in four user-characterization metrics where there is no novice-effect apparent. The change in respect of the session-conformance range is discussed separately later in this Section.

Possible novice-effects

The novice-effect occurs when the longitudinal-developmental change by student-users in the 1998-cohort is an exaggeration of the overall longitudinal-developmental

[3] The mean here is during both study-years, that is 44 sessions.

change. The average session click rate and average session-conformance appear to show the novice-effect since,

$$\text{slope}_{\text{1998-cohort}} > \text{slope}_{\text{overall}} > \text{slope}_{\text{1997-cohort}}$$

for average session click rate and,

$$\text{slope}_{\text{1998-cohort}} < \text{slope}_{\text{overall}} < \text{slope}_{\text{1997-cohort}}$$

for average session-conformance.

Table 6.1 sets out the conditional-regression slopes for these user-characterizations compared by gender and by joint-session-rate as well as by cohort.[4] There is change in all but two of the attribute groups. Neither the 1997-cohort's nor the larger joint-session-rate student-users' average session-conformance shows a change. That is, except for these two groups, the conditional-regression slope is significantly greater or less than one. As discussed above the novice-effect is significant in respect of the average session click rate. There is also a novice-effect for the average session-conformance ($p < .05,\ z = 2.82$).

[4] The conditional distributions are illustrated in Figures 6.1 and E.2 to E.8 in Appendix E.

User-attribute	User-characterization	
	average session click rate	average session-conformance
cohort		
1997-cohort	1.36(0.01)	0.99(0.01)
1998-cohort	1.66(0.01)	0.95(0.01)
gender		
men	1.48(0.01)	0.97(0.01)
women	1.45(0.02)	0.97(0.01)
joint-session-rate		
smaller	1.60(0.01)	0.95(0.01)
larger	1.38(0.01)	1.01(0.01)
Overall	1.46(0.01)	0.97(0.01)

Table 6.1: Cross-tabulation of conditional-regression slopes which appear to show a novice-effect by user-attribute and user-characterization

The conditional-regression slopes are also significantly different for the smaller and larger joint-session-rate student-user's average session click rate ($1.60(0.01) \neq 1.38(0.01)$, $p < .00,\ z = 15$) and average session-conformance ($0.95(0.01) \neq 1.01(0.01)$, $p < .05,\ z = 4.24$). However there is no difference between the gender groups for either user-characterization ($z = 1.34$ and 0 for average session click rate and average session-conformance respectively).

Overall as novice student-users become more experienced so they increase the average number of Web requests (clicks) which they make when locating Web information. This longitudinal-development change is however focussed on student-users with larger joint-session-rates. The conditional-regression slopes in respect of average session click rate for smaller and larger joint-session-rate student-users from the 1998-cohort are 1.58(0.02) and 1.78(0.02) while for the 1997-cohort the slopes are 1.61(0.02) and 1.23(0.02). Hence although the larger joint-session-rate students show a novice-effect the smaller joint-session-rate students do not. The similar analysis by gender indicates that both men and women from the 1998-cohort show a novice-effect since their conditional-regression slopes are 1.73(0.02) and 1.53(0.02) compared with 1.35(0.02) and 1.39(0.02) for the 1997-cohort.

As novice student-users become more experienced and click more they also become more dissimilar in their session Website-vocabulary since they reduce their average session-conformance. However this longitudinal-development change in average session-conformance is focussed in student-users with smaller joint-session-rates since student-users with larger joint-session-rates show no change ($z = 1.00$). The conditional-regression slopes for the different joint-session-rate attribute groups for the 1998-cohort are 0.92(0.01) and 1.00_6(0.01) for the smaller and larger student-users respectively. For the 1997-cohort these slopes are 0.98(0.01) and 1.01_4(0.01) so that slope $_{1998\text{-cohort}}$ is an exaggeration of slope $_{1997\text{-cohort}}$ in both smaller and larger joint-session-rate attribute groups as required but the novice-effect is not significant for the larger group. Hence this novice-effect is more pronounced among the majority of student-users who have a smaller joint-session-rate. The analysis by gender within the cohort groups indicates as previously that both men and women from the 1998-cohort show a novice-effect since their conditional-regression slopes are 0.96(0.01) and 0.95(0.01) compared with 0.98(0.01) and 1.00(0.01) for the 1997-cohort.

The overall conditional-regression slopes for average session click rate and average session-conformance are 1.46(0.01) and 0.97(0.01) respectively. As just discussed, novice student-users change how they locate Web information by increasing their average session click rate and by reducing their session-conformance. Each attribute group increased their average session click rate which suggests that this phenomenon has a structural component but the session-conformance change phenomena is differential in that there is no significant change in respect of the larger joint-session-rate student-user's average session-conformance.

An absence of novice-effects

There are no novice-effects with respect to the other four user-characterization metrics that are being investigated here. In the case of each metric, individuals from the 1998-cohort change no more than individuals from the 1997-cohort.

Comparison of the conditional-regression slopes of user-characterization metrics between student-users associated by different attributes reveals three phenomena.

1. *Stability* over time where the conditional-regression slope is close to one indicating little change between the study-years, for example the average Website-trajectory slope of student-users who have larger than average joint-session-rates ($z = 2.0$),

2. *Similarity* in change in respect of a user-characterization metric between student-users in each of an attribute groups, for example the same increase in aver-

age query-click proportion by both smaller, 1.27(0.01), and larger, 1.28(0.01), student-users ($z = 0.7$),

3. *Difference* in change in respect of a user-characterization metric between student-users in each of an attribute groups, for example the differences in average Website-trajectory slope by student-users in each of the cohorts ($0.86(0.01) \neq 0.96(0.01)$, $p < .00$, $z = 6$).

Three of the overall conditional-regression slopes show significant change, that is the slope is significantly different from one. Overall, there is no change in average Webhost-persistence ($z = 1.0$). The change is greatest in the average query-click proportion for which the overall conditional-regression slope is 1.27(0.01). However there is little or no differential between student-user groups which suggests that like the average session click rate discussed above this increase has a structural component. There is also an overall 12% increase (since the conditional-regression slope is 1.12(0.01)) in average Website-re-request rate. Hence the overall 46% increase in session click rate (see Table 6.1) is partly manifest as greater Website-re-requesting.

Student-users who have larger joint-session-rates change their average Website-re-request rate the least thus their 38% increase in average session click rate (see Table 6.1) must be substantially absorbed within changes in their other user-characterization metrics. It can be seen from Table 6.1 that the student-users who have larger joint-session-rates maintain their average session-conformance ($z = 1.0$) and from Table 6.2 that they reduce least their Website-trajectory slope. Therefore an explanation for these student-users who have larger joint-session-rates also reducing their Webhost-persistence more ($0.98(0.01) < 1.03(0.01), p < .05, z = 3.53$) than the smaller joint-session-rate student-users is that during study-year two their Web information seeking activity continues to be visiting a similar collection of Websites which they visited during study-year one and that they expand their Website-vocabularies by visiting just a few Websites at previously unvisited Webhosts. Thus they depress their Webhost-persistence but maintain most compared to the other student-users, their Website-trajectory slopes.

The overall conditional-regression slope in respect of the Website-trajectory is 0.91(0.01). This means that during study-year two student-users grow their Website-repertoires by only about 91% as much compared to their Website-repertoire growth during study-year one. All the attribute groups of student-user reduce their Website-repertoire growth although those with larger joint-session-rates do so the least (98%). Average Webhost-persistence overall remains the same ($z = 1.0$). Thus, session-by-session, student-users generally visit the same number of Websites from each Webhost during study-year two compared with during study-year one. But half of attribute groups (1998-cohort, men and smaller joint-session-rate) of student-users reduce their

Webhost-persistence while the other half increase their Webhost-persistence. It appears that student-uses with larger joint-session-rates are individually more similar in how they grow their Website-vocabularies during each of the two study-years than are student-users with smaller joint-session-rates. This is because each of the vocabulary characterization metrics shows less change.

Taking the average Webhost-persistence and Website-trajectory slope user characterizations together it appears that men are more individually similar from study-year one to study-year two than are women. In particular women student-users reduce their growth of Website-vocabulary. However since the greatest similarity study-year to study-year is shown by the larger joint-session-rate group of student-users than this gender difference may reflect gender bias in session rate.

The conditional-regression slopes in respect of these three user-characterization metrics analyzed by attribute group are set out in Table 6.2 which also shows the overall conditional-regression slopes.[5]

User-attribute	**User-characterization**			
	average query-click proportion	average Website-re-request rate	average Webhost-persistence	Website-trajectory slope
cohort				
1997-cohort	1.27(0.01)	1.16(0.01)	0.97(0.01)	0.86(0.01)
1998-cohort	1.27(0.01)	1.09(0.01)	1.04(0.01)	0.96(0.01)
gender				
men	1.21(0.01)	1.10(0.01)	1.05(0.01)	0.95(0.01)
women	1.33(0.01)	1.13(0.01)	0.97(0.01)	0.88(0.02)
joint-session-rate				
smaller	1.27(0.01)	1.15(0.01)	1.03(0.01)	0.89(0.01)
larger	1.28(0.01)	1.06(0.01)	0.98(0.01)	0.98(0.01)
Overall	1.27(0.01)	1.12(0.01)	1.01(0.01)	0.91(0.01)

Table 6.2: Cross-tabulation of conditional-regression slopes which do not show a novice-effect by user-attribute and user-characterization

Conditional analysis is used next to investigate how student-users change the variety

[5] The conditional distributions are illustrated in Figures E.9 to E.24 in Appendix E.

of Websites which they visit from session-to-session. This is indicated by change in the range of student-user's session-conformance.

How do student-users change the range of their Web information seeking session-conformance?

The session-conformance metric describes the similarity of a student-user's session Website-vocabulary when compared with all sessions during both study-years (see Chapter three). Sessions which comprise Web requests (or clicks) to the more popular Websites have a larger session-conformance and sessions which are exclusively Web requests to rare Websites (popularity ranking lower than 1000^{th}) have zero session-conformance. The average session-conformance user-characterization facilitates comparing how individual student-users locate Web information but smoothes out the session-to-session variation by individual students-users. How student-users change their average session-conformance is discussed above. Student-users generally reduce their average session-conformance during study-year two compared with study-year one, thus during study-year two student-users resemble each other less in respect of the Websites which they visit. There is a novice-effect which is more pronounced among student-users who have smaller than average joint-session-rates.

The session-conformance range user-characterization for a student-user indicates the extent to which an individual's session-Website-vocabulary replicates itself from session-to-session and hence indicates variety in how an individual locates Web information. If a student-user's session-Website-vocabulary remains the same than so will the session-conformance metric (at whatever value locates the session-Website-vocabulary with respect to overall Website-vocabulary). The
non-longitudinal-developmental investigation of session-conformance range by-user reveals that the frequency distribution for this metric is bi-modal. It is found that the upper and lower parts of the distribution of *range* correspond to whether or not the student-user had visited exclusively rare Websites during *any* session. Those student-users represented in the lower part of the of the range distribution are *conformant* student-users in that they never visit exclusively rare Websites and always locate Web information by visiting at least one or more of the most frequently visited (by-session) Websites. Student-users in the upper part of the distribution have a large range of session-conformance which is connected with their having visited exclusively rare Websites during at least one session. These student-users are referred to as *eclectic*.

Table 4.1 shows that overall more student-users are eclectic during study-year two than during study-year one but this says nothing about how individuals change. The conditional-distribution of the range of student-user's session-conformance examines

how each student-user changes the variety of his session-Website-vocabulary. This distribution, which is illustrated in Figure 6.2, shows that student-users disperse to the four corners of the graph. The overall increase in the range of session-conformance is significant ($p < .00,\ z = 8$). Hence it appears that student-users overall become less self-alike in their Web information seeking which means that there is an increased variety among an individual's sessions while at the same time individuals also become less similar one to another since the average session-conformance reduces. The increased variation among an individual's sessions is thus localized to that individual and Websites are not 'shared'.

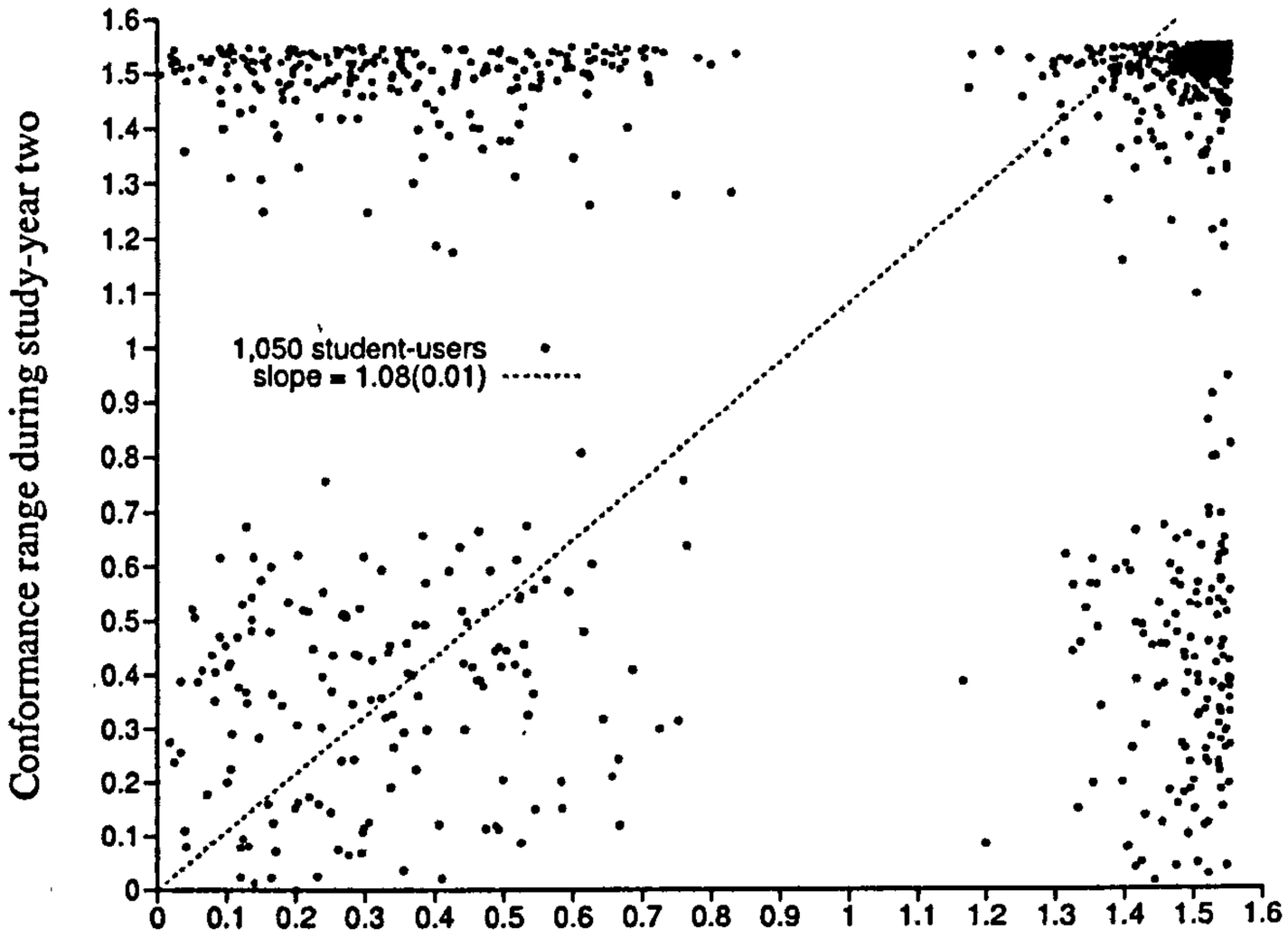

Figure 6.2: Conditional distribution of student-user's conformance range

The 156 points in the lower left quadrant of Figure 6.2 represent student-users who are *double-conformant* that is they eschew rare Websites during each of the study-years. The upper-right quadrant is 517 *double-eclectic* student-users. When analysed separately the double-eclectic student-users show no change over time in their range of session-conformance (slope 1.01(0.01)); the double-conformant student-users increase their range (slope 1.13(0.03), $p < .00,\ z = 4$) indicating a growth in variety of Websites but for these student-users this is constrained to the more popular Websites.

For completeness, the 239 conformant-eclectic student-users (upper left quadrant) have a conditional-regression slope of 4.01(0.01) while the 138 eclectic-conformant student-users (lower right quadrant) have a conditional-regression slope of 0.64(0.01). As expected these slopes are extreme since the conditional-regression line passes through the origin.

The double-conformant, double-eclectic, conformant-eclectic and eclectic-conformant classification provide *conditional-conformance* attributes which partition the student-users. The change effect is analysed nonparametrically using these conditional-conformance attributes to classify student-user frequencies rather than (parametrically) by comparing regression slopes because of the evident clustering of student-users.

The classification of the frequency distribution of the 1997-cohort and 1998-cohort according to student-user's conditional-conformance attributes can be compared using the χ^2 test for consistency. Since $\chi^2 = 5.24 < \chi^2_{3;0.05} = 7.81$, it is accepted that the there is no difference in the conditional-conformance attribute classification between the two cohorts. The test of consistency in respect of gender also shows no difference ($\chi^2 = 2.80 < \chi^2_{3;0.05} = 7.81$). However student-users who have smaller or larger joint-session-rates are different ($p < .05, \chi^2 = 124.31 > \chi^2_{3;0.05} = 7.81$).

Table 6.3 reports student-user frequencies classified according to their conditional-conformance attribute and it can be seen that student-users who have a larger joint-session-rate are mostly double-eclectic. Conversely, student-users who are not double-eclectic mostly have smaller joint-session-rates.

User-attribute	Conditional-conformance attribute			
	double-conformant	conformant-eclectic	double-eclectic	eclectic-conformant
cohort				
1997-cohort	56	87	215	64
1998-cohort	100	152	302	74
gender				
men	84	114	275	67
women	72	125	242	71
joint-session-rate				
smaller	144	185	257	100
larger	12	54	260	38
Overall	156 student-users	239 student-users	517 student-users	138 student-users

Table 6.3: Cross-tabulation of student-user frequency by user-attribute and conditional-conformance attribute

The lack of an inter-cohort difference rules out any novice-effect in respect of the range of session-conformance user-characterization. Equally there are no salient gender differences. However the bias among larger session-rate student-users towards being double-eclectic is quite evident. The strength of the change by larger joint-session-rate student-users towards being eclectic is shown by the 87% ($= \frac{260}{260+38}$) of eclectic and 82% ($= \frac{54}{54+12}$) of conformant student-users during study-year one being eclectic during study-year two. This compares with only 72% ($= \frac{257}{257+100}$) and 56% ($= \frac{185}{185+144}$) of eclectic and conformant smaller joint-session-rate student-users during study-year one being eclectic during study-year two.

Of the seven user-characterizations examined in this Section, the novice-effect is confined to at most the average session click rate and average session-conformance. Thus it appears that proficiency in locating Web information is not associated with any of the other five user-characterizations which describe student-users' Web information seeking activity. Therefore, is there a novice-effect for any of the user-characterization

metrics which describe how student-users use Web information locating services? This question is investigated in the next section.

6.3 How do search-users change their use of Web information location services?

The conditional analyses now examine how 1,002 search-users (out of 1,050 student-users) change their average search-query rate user-characterization and how 584 search-users from the AltaVista-Excite sample change their use of AltaVista and Excite. This is in order to investigate change phenomena associated with how search-users use Web information locating services? The 1,002 student-users are those who used the fourteen principal Web information location services during each of the study-years while the 584 student-users are those who used the AltaVista or Excite Web information location services during each of the study-years (see Chapter five).

The overall regression-slopes for the conditional analyses in respect of all four of the Web information location service user-characterization metrics show significant change, that is each slope is significantly different from one.[6] As above, comparison of the conditional-regression slopes of the user-characterizations between search-users associated by different user-attributes reveals stability, similarity and difference. For example;

> stability over time for men search-user's average search-query count where the slope, (0.99(0.01)) is not significantly different from one ($z = 1.00$).
>
> similarity between student-users in each of the dichotomous attribute groups for example smaller and larger joint-session-rate search-user's average search-query count ($1.16(0.02) \sim 1.11(0.02)$, $z = 1.77$), and,
>
> difference in change, such as men search-users who reduce their search-session proportion count to a greater extent than do women search-users ($0.90(0.02) < 0.99(0.02)$, $p < .05$, $z = 3.18$).

The conditional-regression slopes for each of the Web information location service user-characterizations analysed for each of the user-attribute groups is presented in Table 6.4. 18 out of the 24 attribute groups show a change ($z > 1.00$) about half of which (10 out of 18) is a reduction while the others increase their user-characterization. The average search-query proportion and average search-query

[6] The conditional distributions are illustrated in Figures E.25 to E.40 in Appendix E.

count Web information location service user-characterizations which each overall increase both appear to show a novice-effect.

User-attribute partition	**Web information location service user-characterization**			
	average search-query proportion	search-session proportion	average search-query count	average search-term count
cohort				
1997-cohort	1.00(0.01)	0.88(0.02)	1.06(0.03)	0.92(0.02)
1998-cohort	1.06(0.01)	0.98(0.02)	1.20(0.02)	0.98(0.01)
gender				
men	0.97(0.01)	0.90(0.02)	1.27(0.02)	0.99(0.01)
women	1.11(0.02)	0.99(0.02)	0.91(0.02)	0.92(0.02)
joint-session-rate				
smaller	1.07(0.01)	0.92(0.02)	1.16(0.02)	0.92(0.02)
larger	0.92(0.02)	0.98(0.02)	1.11(0.02)	1.00(0.01)
Overall	1.04(0.01)	0.94(0.02)	1.13(0.02)	0.96(0.01)

Table 6.4: Cross-tabulation of conditional-regression slope by user-attribute and Web information location service user-characterization

As novice search-users become more experienced so they increase their average search-query proportion (that is the proportion of clicks which are search-queries) and their average search-query count (that is the the typical number of search-queries within a search-session).

The longitudinal-development change in average search-query proportion is focussed in women search-users since the conditional-regression slopes for the different gender groups within the 1998-cohort are 0.99(0.02) for men and 1.14(0.02) for women. These compare with 0.92(0.02) and 1.07(0.03) for men and women from the 1997-cohort. Hence although men novices show no change ($z = 0.5$) this is more than offset by the change by women novices. A similar analysis for the joint-session-rate groups within each cohort shows that the larger novices are similar to more experienced

larger search-users in that they *reduce* their average search-query proportion. The novice-effect is confined to smaller novices since the conditional-regression slopes are 1.11(0.02) and 0.92(0.02) for the smaller/larger joint-session-rate groups from the 1998-cohort compared with 1.02(0.02) and 0.93(0.02) for 1997-cohort respectively. The exaggerated increase in average search-query proportion among smaller search-users offsets the lack of any difference between the novice and more experienced search-users who have larger joint-session-rates. The gender difference can again be explained by a bias in the distribution of joint-session-rate.

The longitudinal-developmental change of an increase in average search-query count is also focussed on women (and smaller) novice search-users although as regards the average search-query count all the novice attribute groups show an increase. The different gender groups within the 1998-cohort of search-users have conditional-regression slopes of 1.31(0.03) and 1.04(0.03) for men and women respectively. The equivalent analysis for the 1997-cohort of search-users shows slopes of 1.28(0.03) and 0.78(0.04) so that there is no significant ($z = 0.71$) novice-effect for men novices. The larger novices similarly do not have a novice-effect ($z = 1.60$). As before the overall longitudinal-development change in average search-query count is concentrated in search-users with smaller joint-session-rates. The overall reduction in search-session proportion and average search-term count is ubiquitous in that none of the attribute groups show any increase. Thus searching occurs less (in the sense that more sessions do not include searching) but when searching is undertaken most of the attribute groups of search-users submit more search-queries. (Women generally show a *reduction* in their average search-query count.) Search-users who have larger joint-session-rates are more stable in the sense that they show the least change. The novice-effects with respect to average search-query proportion and average search-query count are concentrated in the search-users who have smaller joint-session-rates.

6.4 Summary and discussion

In Chapter six the longitudinal-developmental investigation uses conditional analysis to examine change in how student-users locate Web information. Conditional analysis compares the characterization of a student-user during study-year two with his (or her) characterization during study-year one. This can help to disambiguate some of the effects of structural change in the Web. For example, it appears that the increase in the proportion of query-clicks which are used has a structural origin. In addition since the analysis is undertaken at the level of the individual it provides a proper basis for interpretation using the notion of a personal Web information infrastructure.

A particular comparison identifies characterizations where the change by novice student-users (that is student-users for whom study-year one is their first year at the institution) is exaggerated compared to the change by non-novice student-users. The direction of change of the characterization metric may be either positive or negative. This is described as the novice-effect and is equivalent to there being a learning-curve associated with the characteristic in question. The presence of a novice-effect is a necessary condition for the characteristic to be associated with proficiency in locating Web information. The notion of proficiency is left undefined but some general properties are taken as being understood; not all student-users are equally proficient at locating Web information, proficiency increases with experience, the rate of increase of proficiency reduces as proficiency increases, and there are observable indicators of proficiency. Hence if a characteristic does not show the novice-effect then the characteristic does not indicate proficiency with respect to locating Web information.

Only two of the user-characterization metrics and two of the Web information location service user-characterization metrics show a novice-effect. These are firstly the student-user's average session click rate and average session-conformance, and secondly the average search-query proportion and average search-query count. That is, novice student-users change how they locate Web information by increasing their average session click rate and reducing their session-conformance, and novice search-users increase their average search-query proportion and their average search-query count. None of the other five user-characterizations or two Web information location service user-characterizations are therefore associated with proficiency in locating Web information.

Both average session click rate and average session-conformance are possible indicators of proficiency. For example proficiency may be indicated inversely by the average session-conformance since it is seen that,

> not all student-users have the same average session-conformance,
>
> average session-conformance reduces with experience, and
>
> the rate of reduction of average session-conformance reduces as the reduction increases.

Change in how student-users locate Web information is not uniform between different groups as defined by their gender, joint-session-rate and conformance attributes. Women change more than men student-users and student-users with smaller joint-session-rates change more than do student-users with larger joint-session-rates. For example the novice-effect for average session click rate is more pronounced among

women student-users while the novice-effect for average session-conformance is more pronounced among student-users with a smaller joint-session-rate. In particular while the smaller joint-session-rate search-users show a novice-effect, the larger joint-session-rate search-users do not. Student-users with larger joint-session-rates are generally more stable (show least change) in how they locate Web information.

It is also found that how men student-users grow their Website-vocabularies during each of the two study-years is more similar than is how women grow their Website-vocabularies and that compared to student-users with smaller joint-session-rates, those with larger joint-session-rates continue during study-year two to visit a similar collection of Websites to the Websites which they visited during study-year one. In addition it appears that student-users who have larger joint-session-rates tend to visit only a few Websites at previously unvisited Webhosts.

Overall, student-users change to become less similar one to another in their Web information seeking. Individual's sessions also become less alike but become more distinctive of the individual. That is, the collection of Websites which each individual visits during each session becomes more varied while at the same time the difference between the collections of Websites of different individuals becomes greater. This change towards greater eclecticism is strongest among student-users with larger joint-session-rates. There is no gender difference.

The finding of both more stability in the individual Website vocabularies of student-users with larger joint-session-rates and that these vocabularies become more individually distinctive lends weight to an interpretation whereby student-users respond to their personal information needs by developing a distinctive vocabulary of Websites or territory which increasingly can satisfy the information task which they have in hand.

The investigation into change in how student-users use Web information location services ('search-engines') is based on the AltaVista-Excite sample, see Chapter five. When analysed by individual it is found that student-user's 'searching' reduces although when sessions do include searching, more search-queries are submitted. However this metric, the average number of search-queries submitted during a session, is viewed cautiously since it may depend on the duration of the session which is not known. The number of terms in each search-query submitted is smaller during study-year two compared with study-year one. This reduction applies to all the attribute groups except the larger joint-session-rate search-users among whom there is no change in average search-term count.

Following the individually based analysis of this Chapter then the findings can be more properly interpreted within the framework of a personal Web infrastructure. The territory or collection of Website visiting and revisiting developed by each

student-user is interpreted as being the Web information infrastructure which is exploited by each individual to resolve their information tasks. Over time as the personal Web information infrastructure or territory becomes more strongly developed so the student-user has less need to stray outside of it. This territorial self-reliance is reflected also in the reduction in 'searching' as student-users' territory becomes more strongly developed.

Conformant student-users have less distinctive territories however these still appear to satisfy the user's information needs. Over time conformant student-users tend to become eclectic which is interpreted as them building their personal Web information infrastructure.

The development of a personal Web information infrastructure can also be seen as corresponding to an individual's proficiency in locating Web information since both are inversely related to the average session-conformance.

7

Student-users are 'territorial' in how they locate Web information

7.1 Introduction

The goal of the research is to describe what student-users do when they use the Web. Key requirements of this description are that it be a description of users (not of sessions or of the Web) and that the description should be reproducible. It would also be helpful if the description distinguishes between users.

A motivation for setting this goal is the current absence from the literature of any equivalent work. However it is previously identified that studying what Web users do requires both closely defined metrics to characterize their Web information seeking actions and analyses which are problem centered and consider individuals' real world Web information seeking spanning multiple sessions.

The large scale two year long observation of student-user's Web information seeking activity (operationalised as their clicking or making Web requests) achieves this and facilitates both longitudinal-developmental analyses and repeat study analyses. These help to disambiguate changes in a student-user's Web information seeking activity due to his (or her) increased proficiency in locating Web information and those changes which may be caused by structural change in the Web.

The study takes the form of an interpretive analysis of (client-side) Web logs which capture each Website visited or 'clicked on' by each student during each daily session. Seven user-characterizations were found to be useful in discovering what it is that users do and which address the aim of the research, 'How do students who use the Web locate information resources'?

A geographical (or spatial) metaphor is often used to describe the Web and its use. Developing this metaphor, the collection of visits and revisits to Websites by a user is described here as the user's *territory*. (A student-user's Website vocabulary is the Websites included in the territory; the notion of territory entails also the proportions of visits to different Websites in the vocabulary.)

The thesis presented here is that student-users are 'territorial' in how they locate Web information. By this I mean that over time student-users' territories become characteristic of the individual (or more precisely, of the Web information needs of the individual) and that student-users increasingly satisfy their Web information needs from within their own territories.

The thesis of territoriality is interpreted as student-users developing their personal Web information infrastructures. From an external perspective the Web may be regarded as a single homogeneous information environment however the findings of this investigation suggest that each student-user constructs a personalised Web information environment which corresponds to a personal Web information infrastructure. The implications which arise from this suggestion are discussed below.

Equating 'searching' with a user submitting a search-query-click to a 'search-engine' and 'surfing' with a user submitting other clicks then the overall ratio of 'searching' to 'surfing' remains at about 12% during each study-year. But 'searching' is not uniform either by-session or by-user. In particular 'search-engines' are only used in slightly less then half of all sessions (although over a sufficiently long period of time all student-users make some use of a 'search-engine'). Over time it also appears that student-users reduce their reliance on 'search-engines'. In addition student-users who use the Web more often and student-users who are eclectic are less likely to use 'search-engines' when compared to those student-users who use the Web less often or who are conformant. 'Searching' therefore appears to be just a means by which student-users develop their personal Web information infrastructure. As this infrastructure becomes stronger so it appears the need to 'search' diminishes.

This concluding Chapter has three principal Sections which are **Summary of findings**, **Strengths and weaknesses of the investigation** and **Implications and further work** together with a final **Conclusion.**

7.2 Summary of findings

The seven user-characterizations which are used in the analysis are (i) average session click rate, (ii) average query-click proportion, (iii) average Website-re-request rate,

(iv) average Webhost-persistence, (v) Website-trajectory slope, (vi) average session-conformance, and (vii) session-conformance range. Conclusions are refined by considering the similarities and differences between and within groups of student-users. These groups are defined according to three user-attributes (gender, session-rate and conformance).

Contrary to what is possibly a popular belief, it was discovered that most student-users used the Web to locate Web information on only a few occasions and that, on those occasions when they did use the Web they used it only a little. A consequence of this is that the mean values of user-characterizations overstate what typical student-users do. This is even more the case here since the research design excludes the most infrequent users of the Web. However since using email (for example) was also excluded because it is not a Web information resource, then an impression of extensive Web use by students may be a reflection of extensive email use by students. Also contrary to what may be popularly believed, there is no gender difference in how students who use the Web locate information resources.

Because most student-users most of the time make little use of the the Web to locate information resources, the distributions of user-characterizations are constrained and distorted. In consequence nonparametric statistical techniques are used to analyse the user-characterizations.

The student-users investigated are not homogeneous in their user-characterizations and no single user-characterization in isolation adequately describes how they use the Web to locate information resources. This is because of the large variation among student-users in what they do and also because of how much each student-user varies over time in what they do. Differentiating between how Web users locate information resources requires multiple user-characterizations, and especially user-characterizations which consider multiple sessions.

A particular area of research interest within library and information science is centred on differentiating the information seeking of 'experts' and 'non-experts'. Without attempting a definition of 'expert' in the context of Web information seeking, the general notion of *proficiency* and its associated rate of change is used to investigate the changes in individual student-user's user-characterizations. It is concluded that only the average session-conformance user-characterization can be a reliable indicator of proficiency. The average session-conformance user-characterization is interpreted also as an indicator of the development of a personal Web information infrastructure. It appears that the frequency of use of the Web is not an indicator of proficiency and so in the absence of other evidence, users who use the Web more frequently should not be taken as being any more expert than users who make less frequent use of the Web.

Structural change in the Web is a confounding factor. It appears that the increase in query-clicking is structural and is due to an increased exploitation of this technique by Website designers. Hence a description of how users locate Web information based on query-clicking would be describing the Web. Both average session click rate and average Website-re-request rate also appear to be either structural or associated with a possible change in the duration of student-users' sessions. The duration of sessions cannot be determined because only the time of the *request* for Web information is known. However the average query count to Web 'search-engines' remains the same during each study-year. Since this count would also be sensitive to a change in session duration then it suggests that average session duration remains the same.

Student-users overall, regardless of gender, change to become more self-alike in their Web information seeking and less similar one to another. Hence, the territory which each individual visits during each session becomes less varied while the difference between individuals' territories becomes greater. The change towards greater eclecticism or distinctiveness in Web information seeking is stronger among student-users with larger joint-session-rates.

The conclusion that Web users are distinctive in respect of their territories is seen also in the "surprising lack of overlap" reported by Cockburn & McKenzie (2001, p. 917) The extent of 'self-alikeness' may be inferred from Kelly's calculation (2002, p. 362) that 73% of users' Web requests would be served from a sufficiently large local browser cache and 58% of the remaining Web requests could be served from a shared proxy cache. Hence 73% of each user's Web requests are self-alike or territorial and about 11% of each user's Web requests are one-off and not shared with other users. One-off requests which are shared with another user amount to only about 16% of users' Web requests which corresponds to the trend towards greater individual eclecticism among the student-user found here.

The development by student-users of a personal Web information infrastructure is hypothesized as an explanation for student-users changing to become more distinctive and less similar one to another. This development is associated with student-users constructing personalized Web information environments so that, increasingly, Web resources requested by student-users are already located within their personalized Web information environment and hence less use is made of 'searching'.

Comparison of the analysis of search-queries from the student-users with published analyses shows a good level of agreement which suggests that the student-users' Web information seeking is similar to users' Web information seeking more generally. In addition, the comprehensive inclusion of student-users in the study without any form of explicit selection supports the contention that the students studied are a validly representative cross section of Web users within the age group.

7.3 Strengths and weaknesses of the investigation

The investigation is a valid study of what it is that a large group of Web users do in the sense of it being an unobtrusive extended observation of student-users at a UK higher education institution during two academic years. The observation takes the form of a client-side Web log which reliably captures all the Websites visited by each student-user. The investigation is therefore a study of real world Web information seeking activity. A particular strength is that the 1,050 student-users represented in the Web log provide a complete survey of all the student-users from two cohorts of students.

The investigation focuses on how Web information resources are located by considering *Web information* only which definition excludes, for example, using email. This contributes to the interpretation of the Web log being reliably founded on an explicit and rigorous user-focussed analysis of Web information seeking activity which uses closely defined metrics. The analysis design takes the form of both a longitudinal analysis and a repeat study which identifies both reproducible results and the effects of change. Particular attention is paid to the possible effects of a structural change in the Web and of individual change which might be associated with proficiency in using the Web.

The investigation thus focuses on student-users' observable Web information seeking actions and the findings report what it is that student-users do. It thus fills an omission in the existing literature.

The study makes no attempt to understand student-users' information problems, nor the extent to which the Web information resources located by student-users satisfy their information needs. The research assumes that the information problems which are resolved by student-users are representative by virtue of the large scale real world nature of the investigation.

Whether or not the research findings can be generalized either to a broader student population in the first instance or to a broader population of Web information seekers theoretically depends (assuming a task and individual-difference model of information behaviour) on both how representative are the information problems and whether or not student-users information seeking in respect of given tasks is similar to other Web users. Both of these are open questions. Comparison with the Excite studies suggests a similarity in information seeking since the studies are in agreement in respect of the reduction in search-terms per search-query. The student-users here represent those attending the institution rather than to just computer/information science specialists as is frequently the case with information behaviour research. This combined with the large scale real world nature of the investigation suggests that

the information problems can be taken as being representative. Therefore there is a better basis for generalizing these findings than those of other studies. However extrapolation of the findings must be cautious since because of the infancy of Web information behaviour research we do not know the conditions under which extrapolation may be reasonable.

The 'territorial' thesis, including the construction of a personalised Web information environment, as a description of how student-users locate Web information does not depend exclusively on analyses based on using the session-conformance metric however these analyses do provide substantial support. The session-conformance metric is a novel approach for interpreting user Web logs and requires study in its own right. In particular the way that it was used here is dictated by the need to discriminate between sessions. The cutoff level of the top 1,000 Websites is thus pragmatic but it is also arbitrary. Further work is needed to understand how sensitive any findings may be to changing this cutoff level.

The session-conformance metric is also lexicographic as regards the Website url-string. There is no semantic or other investigation of the substantial content of the Website. Therefore there is no means of knowing within this investigation how similar two Websites may be as regards their information content or meaning. Hence it could be argued that the territories of two student-users are semantically identical even though the the Websites are lexicographically different. Issues of semantic and content similarity are discussed below.

The rate of structural change in the Web and its detailed effects on particular information seeking metrics is unknown. It is certain that detailed features concerning Web information seeking will change even within say two years and hence the values found for the metrics are temporally limited. However many of the detailed results here show that the characterization metrics are small and are approaching a minimum values. For example, host-persistence is reducing but, by definition, cannot reduce smaller then one Website per Webhost. In addition, as the Web gets bigger so the the effect of any particular structural change is proportionately smaller.

Thus, although the territorial thesis of how student-users locate Web information is related to the Web as found during the study, nevertheless the nature of the thesis makes it robust compared to a specific functional description of information seeking. For example, a description predicated on particular types of information resource such as *mpeg* audio-visual files or html hypertext files.would be overtaken by file sharing services (such as Napster) being outlawed and more flexible Web presentation standards such as xml replacing html.

7.4 Implications and suggestions for further work

The thesis of territoriality may be explained as users constructing personalized Web information environments which they exploit to satisfy their individual Web information needs. This explanation is attractive since it offers an interpretation of Web information behaviour which is consistent with Marchionini's notion of a "*personal information infrastructure*". He constructs this notion in order to describe "an individual person's collection of abilities, experience, and resources to gather, use, and communicate information". The "level of development of a person's information infrastructure is roughly analogous to the level of his or her information literacy" (Marchionini, 1995, p. 11).

It is hypothesized that Web users develop a *personal Web information infrastructure* which supports each user to resolve his (or her) Web information problems. As users strengthen their personal Web information infrastructure so they become more adept at resolving their individual Web information problems.

More proficient Web users are thus only more proficient in the limited sense of the strength of their own personal Web information infrastructures, or using Web information resources to resolve their own Web information problems. Because of territoriality, a user's *general* Web information literacy will develop progressively less than his (or her) *personal* Web information literacy. Hence the implication is that more proficient users collectively have a mix of heterogenous and distinct personalized Web information environments only one of which is known to each user. Therefore the Web is not an homogeneous information resource about which proficient users all have equivalent knowledge (of how to locate Web information resources).

Thus outside of his (or her) personal Web information environment an experienced Web user is no more Web literate than a novice Web user. Studies of Web information behaviour which fail to take this into account may thus have difficulty distinguishing users who have differing Web experience.

An area of future work is to understand the nature of personal Web information environments, for example in order to discover how much of an overlap in Web-sites there may be at a semantic level and how personal Web information literacy may be supported. The profile analysis of Web information seeking sessions mentioned in Chapter four provides a potential technique for investigating territories, how personal Web information environments are developed and how users with different approaches to developing their personal Web information environments may be distinguished.

Not all users are eclectic or develop a distinctive territoriality. Some, but a reducing minority, of users are conformant and when locating Web information always visit

one or more of the most frequently visited Websites. A second area for future work is thus to investigate why this should be so and what it is that results in this difference in how users locate Web information.

The identification of territoriality by Web users makes use of the session-conformance metric which is a novel approach for interpreting user Web logs. This requires study in its own right. Two aspects are of interest. Firstly the metric is based on the lexicographic similarity of users' territory (or collection of Websites) rather than in similarities of content or semantic similarity between Website collections. Therefore what is the effect of different bases of session-conformance? Secondly the session-conformance metric uses an arbitrary cutoff level of the top 1,000 Websites to differentiate conformant and eclectic student-uses. What is the sensitivity of the metric to changing the cutoff level?

The session-conformance also makes full use of the frequencies of a user visiting different Websites. This information is not normally available outside of comprehensive client-side based studies as undertaken here. However proxy-servers are informed only of the *different* Websites. Therefore a final suggestion for further work is to investigate the territoriality of other groups of users based on proxy-server logs and a modified session-conformance metric.

7.5 Conclusion

Student-users are territorial in how they locate their information resources. That is, each individual has a Web territory made up of visits and revisits to Websites which becomes more distinctive over time and is characteristic of the Web information needs of the student-user. Student-users become more territorial over time in that they increasingly locate Web information resources from within their own territories. More proficient student-users are more territorial and rely less on 'searching' to locate Web resources.

Territoriality can be interpreted as student-users developing their personal Web information infrastructure and constructing personalized Web information environments.

Territoriality in locating Web information resources represents a users' perspective of the Web. This is completely opposite to the perspective of the Web taken by the major 'search-engines' which crawl the Web using robots. Hence studying users' territories may provide a way to improve the relevance of a resource to an information problem when seeking Web information. The resource's authority would be improved by it being included in the territory of another user who has similar information problems rather than the resource authority being based on a Web graph (Brin &

Page, 1998; Kleinberg, 1999) without knowing whether or not users actually refer to it.

References

Aalbersberg, I. J. (1994). A document retrieval model based on term frequency ranks. In Croft, W. B. & van Rijsbergen, C. J., eds., *Proceedings of seventeenth annual international ACM SIGIR conference on research and development in information retrieval, Dublin, July 3–6, 1994*, pp. 163–172, Association for Computing Machinery SIGIR, London: Springer.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Harlow, England: ACM Press / Addison Wesley.

Bailey, P., Craswell, N. & Hawking, D. (1999). Chart of darkness: mapping a large intranet. [Online], available from: <url:http://pigfish.vic.cmis.csiro.au/~nickc/pubs/cod.ps.gz>, [accessed 30 May 2002].

Barford, P., Bestavros, A., Bradley, A. & Crovella, M. E. (1999). Changes in Web client access patterns: characteristics and caching implications. *World Wide Web*, 2(2), pp. 15–28.

BBC (2001). UK Web stats "notoriously inaccurate". *BBC News: Sci/Tech*, [online], available from: <url:http://news.bbc.co.uk/hi/english/sci/tech/newsid_1422000/1422275.stm>, [accessed 22 April 2002].

Bell, W. J. (1990). *Searching behaviour: the behavioural ecology of finding resources*. London: Chapman and Hall.

Berners-Lee, T. (1999). *Weaving the Web: the past, present and future of the World Wide Web by its inventor*. London: Texere.

Berners-Lee, T., Masinter, L. & McCahill, M. (1994). RFC 1738: Uniform resource locators (URL). [Online], available from: <url:http://www.w3.org/Addressing/rfc1738.txt>, [accessed 24 October 2000].

Bharat, K., Broder, A., Dean, J. & Henzinger, M. R. (2000). A comparison of techniques to find mirrored hosts on the WWW. *Journal of the American Society for Information Science*, 51(12), pp. 1114–1122.

Bijleveld, C. C. J., van der Kamp, L. J. T., Mooijaart, A., van der Kloot, W. A., van der Leeden, R. & van der Burg, E. (1998). *Longitudinal data analysis: designs models and methods*. London: Sage.

Bilal, D. (2000). Children's use of Yahooligans! Web search engine: I: cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for Information Science*, 51(7), pp. 646–665.

Bilal, D. (2001). Children's use of Yahooligans! Web search engine: II: cognitive and physical behaviors on research tasks. *Journal of the American Society for Information Science*, 52(2), pp. 118–136.

Bilal, D. & Kirb, J. (2002). Differences and similarities in information seeking: children and adults as Web user. *Information processing and management*, 38(5), pp. 649–670.

Borgman, C. L. (1986). Why are online catalogs hard to use: lessons learned from information-retrieval studies. *Journal of the American Society for Information Science*, 37(6), pp. 387–400.

Borgman, C. L. (2000). *From Gutenberg to the global information infrastructure: access to information in the networked world*. London: MIT Press.

Bowers, N. & Taylor, R. (2000). Netscape:History – object class for accessing Netscape history database. [Online], available from: <url:http://search.cpan.org/doc/NEILB/Netscape-History-3.01/Netscape/History.pm>, [accessed 29 May 2002].

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1–7), pp. 107–117.

Brown, A. & Dowling, P. (1998). *Doing research/reading research: a mode of interrogation for education*. London: Falmer Press.

Burton, M. C. & Walther, J. B. (2001). A survey of Web log data and their application in use–based design. In *Proceedings of the 34th Hawaii international conference on system sciences, Maui, Hawaii, January 3–6, 2001*, IEEE Computer Society, Los Alamitos: IEEE, [online], available from: <url:http://dlib.computer.org/conferen/hicss/0981/pdf/09815023.pdf>, [accessed 11 June 2001].

Carroll, J. B. (1999). Expert Internet information access. *Journal of educational computing research*, 20(3), pp. 209–222.

Catledge, L. D. & Pitkow, J. E. (1995). Characterizing browsing strategies in the World Wide Web. *Computer networks and ISDN systems*, 27(6), pp. 1065–1073.

Chalmers, R. (2000, 11 November). Surf like a bushman. *New Scientist*, 168(2264), pp. 39–41.

Chen, B., Wang, H., Proctor, R. W. & Salvendy, G. (1997). A human-centered approach for designing World-Wide Web browsers. *Behavior research methods, instruments and computers*, 29(2), pp. 172–179.

Chen, C., Czerwinski, M. & Macredie, R. (2000). Individual differences in virtual environments: introduction and overview. *Journal of the American Society for Information Science*, 51(6), pp. 499–507.

Chen, H. & Cooper, M. D. (2001). Using clustering techniques to detect usage patters in a Web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11), pp. 888–904.

Cheung, D. W., Kao, B. & Lee, J. (1998). Discovering user access patterns on the World Wide Web. *Knowledge based systems*, 10(7), pp. 463–470.

Choo, C. W., Detlor, B. & Turnbull, D. (1998). A behavioral model of information seeking on the Web: preliminary results of a study of how managers and IT specialists use the Web. In Preston, C. M., ed., *Proceedings of the 61st ASIS annual meeting, Pittsburg, Pennsylvania, October 24–29, 1998*, vol. 35, pp. 290–302, American Society for Information Science, Medford, New Jersey: Information Today for ASIS, [also online], available from: <url:http://choo.fis.utoronto.ca/fis/respub/asis98/>, [accessed 25 January 2001].

Christ, M., Krishnan, R., Nagin, D., Kraut, R. & Günther, O. (2001). Trajectories of individual WWW usage: implications for electronic commerce. In *Proceedings of the 34th Hawaii international conference on system sciences, Maui, Hawaii, January 3–6, 2001*, pp. 2794–2802, IEEE Computer Society, Los Alamitos: IEEE, [also online], available from: <url:http://www.computer.org/>, [accessed 2 June 2001].

Christiansen, T. & Torkington, N. (1998). *Perl cookbook*. Cambridge, MA: O'Reilly.

Cockburn, A. & Jones, S. (1996). Which way now: analysing and easing inadequacies in WWW navigation. *International journal of human–computer studies*, 45(1), pp. 105–129.

Cockburn, A. & McKenzie, B. (2001). What do Web users do: an empirical analysis of Web use. *International journal of human–computer studies*, 54(6), pp. 903–922.

Cool, C. & Spink, A. (2002). Issues of context in information retrieval (IR): an introduction to the special issue. *Information processing and management*, 38(5), pp. 605–611.

Cooley, R., Mobasher, B. & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. *Knowledge and information systems*, 1(1), pp. 5–32.

Cooper, M. D. (1998). Design considerations in instrumenting and monitoring Web-based information retrieval systems. *Journal of the American Society for Information Science*, 49(10), pp. 903–919.

Cooper, M. D. (2001). Usage patterns of a Web-based library catalog. *Journal of the American Society for Information Science and Technology*, 52(2), pp. 137–148.

Cooperative Association for Internet Data Analysis (2002). The CAIDA Web site: CAIDA Home. [Online], available from: <url:http://www.caida.org/>, [accessed 5 July 2002].

Cothey, V. (2001). Letter: a comparison of techniques to find mirrored hosts on the WWW. *Journal of the American Society for Information Science and Technology*, 52(8), pp. 601–602.

Cothey, V. (2002). A longitudinal study of World Wide Web users' information searching behavior. *Journal of the American Society for Information Science and Technology*, 53(2), pp. 67–78.

Cothey, V. J. (1998). *Information retrieval from the World Wide Web: an investigation into users' searching activities*. Master's dissertation, University of Bristol, Bristol.

Cove, J. F. & Walsh, B. C. (1987). Browsing as a means of online text retrieval. *Information services and use*, 7(6), pp. 183–188.

Cove, J. F. & Walsh, B. C. (1988). Online text retrieval via browsing. *Information processing and management*, 24(1), pp. 31–37.

Cunha, C. R., Bestavros, A. & Crovella, M. E. (1995). Characteristics of WWW client-based traces. Technical report BU-CS-95-010, Computer Science Department, Boston University, [online], available from: <url:http://www.cs.bu.edn/techreports/95-010-www-client-traces.ps.Z>, [accessed 24 October 2000].

Cunha, C. R. & Jaccoud, C. F. B. (1997). Determining WWW user's next access and its application to pre-fetching. Technical report 97-004, Computer Science Department, Boston University, [online], available from: <url:http://www.cs.bu.edn/techreports/97-004-userbehaviorprediction.ps.Z>, [accessed 24 October 2000].

Denzin, N. K. (1978). *The research act: a theoretical introduction to sociological methods*. London: McGraw-Hill, second edition.

Ellis, D. (1992). The physical and cognitive paradigms in information retrieval research. *Journal of documentation*, 48(1), pp. 45–64.

Fidel, R., Davies, R. K., Douglass, M. H., Holder, J. K., Hopkins, C. J., Kushner, E. J., Miyagishima, B. K. & Toney, C. D. (1998). A visit to the information mall: Web searching behavior of high school students. *Journal of the American Society for Information Science*, 50(1), pp. 24–37.

Ford, N., Miller, D. & Moss, N. (2001). The role of individual differences in Internet searching: an empirical study. *Journal of the American Society for Information Science and Technology*, 52(12), pp. 1049–1066.

Ford, N., Miller, D. & Moss, N. (2002). Web search strategies and retrieval effectiveness: an empirical study. *Journal of documentation*, 58(1), pp. 30–48.

Ford, N., Wilson, T., Foster, A., Ellis, D. & Spink, A. (2000). Individual differences in information seeking: an empirical study. In Kraft, D. H., ed., *Proceedings of the 63rd ASIS annual meeting, Chicago, Illinois, November 12–16, 2000*, vol. 37, pp. 14–24, American Society for Information Science, Medford, New Jersey: Information Today for ASIS.

Frants, V. I., Kamenoff, N. I. & Shapiro, J. (1993). One approach to classification of users and automatic clustering of documents. *Information processing and management*, 29(2), pp. 187–195.

Fu, Y., Sandhu, K. & Shih, M. (1999). Clustering of Web users based on access patterns. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, California, August 15–18,*

1999: workshop on Web usage analysis and user profiling, August 15, 1999, ACM SIGKDD, New York: Association for Computing Machinery, [also online], available from: <url:http://www.acm.org/sigkdd/proceedings/webkdd99/papers/fu.htm>, [accessed 26 January 2000].

Georgia Tech Research Corporation (1999). GVU's WWW user surveys. [Online], available from: <url:http://www.gvu.gatech.edu/user_surveys/>, [accessed 1 July 2002].

Goldstein, H. (1979). *The design and analysis of longitudinal studies: their role in the measurement of change*. London: Academic Press.

Goodrum, A. & Spink, A. (2001). Image searching on the Excite Web search engine. *Information processing and management*, 37(2), pp. 295–311.

Graham-Cumming, J. (1997). Hits and miss-es: a year watching the Web. *Computer networks and ISDN systems*, 29(8–13), pp. 1357–1365.

Graham-Cumming, J. (1998). Optimal Internet monitor quotes and anecdotes, private communication.

Greenberg, S. (1993). *The computer user as toolsmith: the use, reuse, and organization of computer-based tools*. Cambridge: Cambridge University Press.

Gross, M. (1995). The imposed query. *RQ*, 35(2), pp. 236–243.

Gross, M. (1998). The imposed query: implications for library service evaluation. *Reference and user services quarterly*, 37(3), pp. 290–299.

Gross, M. (1999). Imposed queries in the school library media center: a descriptive study. *Library and information science research*, 21(4), pp. 501–521.

Gross, M. (2001). Imposed information seeking in public libraries and school media centers: a common behaviour? *Information research*, 6(2), [online], available from: <url:http://www.shef.ac.uk/~is/publications/infres/6-2/paper100.html>, [accessed 22 January 2001].

He, D. & Göker, A. (2000). Detecting session boundaries from Web user logs. In *Proceedings of the 22nd annual colloquium on information retrieval research, Cambridge, April 5–7, 2000*, Information retrieval specialist group of the British Computer Society, [online], available from: <url:http://irsg.eu.org/irsg2000online/papers/source/he.pdf>, [accessed 31 January 2001].

He, D., Göker, A. & Harper, D. J. (2002). Combining evidence for automatic Web session identification. *Information processing and management*, 38(5), pp. 727–742.

Heaps, H. S. (1978). *Information retrieval: computational and theoretical aspects*. London: Academic Press.

Higher Education Funding Council for England (1997). *Profiles of higher education institutions: 1997*. Bristol: Higher Education Funding Council for England.

Higher Education Statistics Agency (1999). *Students in higher education institutions: 1997/8*. Cheltenham: Higher Education Statistics Agency.

Higher Education Statistics Agency (2000). *Students in higher education institutions: 1998/9*. Cheltenham: Higher Education Statistics Agency.

Higher Education Statistics Agency (2001). *Students in higher education institutions: 1999/2000*. Cheltenham: Higher Education Statistics Agency.

Hoelscher, C. & Strube, G. (1999). Searching on the Web: two types of expertise. In Hearst, M., Gey, F. & Tong, R., eds., *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, California, August 15–19, 1999*, pp. 305–306, ACM SIGIR, New York: Association for Computing Machinery.

Hölscher, C. & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer networks*, 33(1–6), pp. 337–346.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. London: Addison-Wesley.

Hsieh-Yee, I. (2001). Research on Web search behavior. *Library and information science research*, 23(2), pp. 167–185.

Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. & Lukose, R. M. (1998). Strong regularities in World Wide Web surfing. *Science*, 280(5360), pp. 95–97, [3 April 1998].

Ingwersen, P. (2001). Cognitive information retrieval. In Williams, M. E., ed., *Annual review of information science and technology*, vol. 34, 1999/2000, chap. 1, pp. 3–52, Medford, NJ: Information Today for American Society for Information Science & Technology.

Jansen, B. J. (2000a). The effect of query complexity on Web searching results. *Information research*, 6(1), [online], available from: <url:http://www.shef.ac.uk/~is/publications/infres/paper87.html>, [accessed 22 January 2001].

Jansen, B. J. (2000b). An investigation into the use of simple queries on Web IR systems. [Online], available from: <url:http://jimjansen.tripod.com/academic/pubs/ir2000/ir2000.html>, [accessed 29 January 2001].

Jansen, B. J. (2000c). Web searchers, they're smarter than they first appear. [Online], available from: <url:http://jimjansen.tripod.com/academic/pubs/ir2000/forum2000.html>, [accessed 29 January 2001].

Jansen, B. J., Goodrum, A. & Spink, A. (2000a). Searching for multimedia: video, audio, and image Web queries. *World Wide Web*, 3(4), pp. 249–254.

Jansen, B. J. & Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science*, 52(3), pp. 235–246.

Jansen, B. J. & Spink, A. (2000). Methodological approach in discovering user search patterns through Web log analysis. *Bulletin of the American Society for Information Science*, 27(1), [also online], available from: <url:http://www.asis.org/Bulletin/Oct-00/janses_spink.html>, [accessed 8 January 2001].

Jansen, B. J., Spink, A., Bateman, J. & Saracevic, T. (1998a). Real life information retrieval: a study of user queries on the Web. *SIGIR forum*, 32(1), pp. 5–17.

Jansen, B. J., Spink, A., Bateman, J. & Saracevic, T. (1998b). Searchers, the subjects they search, and sufficiency: a study of a large sample of Excite searches. In *Proceedings of the world conference of the WWW, Internet and intranet, Orlando, Florida, November 7–12, 1998*, Association for the Advancement of Computing in Education, [online], available from: <url:http://jimjansen.tripod.com/academic/pubs/webnet98.pdf>, [accessed 29 January 2001].

Jansen, B. J., Spink, A. & Pfaff, A. (2000b). Linguistic aspects of Web queries. In Kraft, D. H., ed., *Proceedings of the 63rd ASIS annual meeting, Chicago, Illinois, November 12–16, 2000*, vol. 37, pp. 169–176, American Society for Information Science, Medford, New Jersey: Information Today for ASIS.

Jansen, B. J., Spink, A. & Pfaff, A. (2000c). Web query structure: implications for IR system design. In *Proceedings of the fourth world multiconference on systemics, cybernetics and informatics, Orlando, Florida, July 23–26, 2000*, vol. 2: information systems development, International Institute of Information and Systemics, [also online], available from: <url:http://jimjamsen.tripod.com/academic/pubs/sci2000/sci2000.pdf>, [accessed 15 June 2001].

Jansen, B. J., Spink, A. & Saracevic, T. (1998c). Failure analysis in query construction: data and analysis from a large sample of Web queries. In *Proceedings of the third ACM conference on digital libraries, Pittsburgh, Pennsylvania, June 23–26, 1998*, pp. 289–290, ACM Digital Libraries, New York: Association for Computing Machinery.

Jansen, B. J., Spink, A. & Saracevic, T. (1999). The use of relevance feedback on the Web: implications for Web IR system design. In *Proceedings of the world conference of the WWW, Internet and intranet, Honolulu, Hawaii, October 24–30, 1999*, Association for the Advancement of Computing in Education, [online], available from: <url:http://jimjamsen.tripod.com/academic/pubs/webnet99.pdf>, [accessed 25 January 2001].

Jansen, B. J., Spink, A. & Saracevic, T. (2000d). Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information processing and management*, 36(2), pp. 207–227.

Jones, S., ed. (1999). *Doing Internet research: critical issues and methods for examining the net*. London: Sage.

Jones, S., Gatford, M., Do, T. & Walker, S. (1997). Transaction logging. *Journal of documentation*, 53(1), pp. 35–50.

Kanji, G. K. (1999). *100 statistical tests*. London: Sage, new edition.

Kelly, T. (2002). Thin-client Web access patterns: measurements from a cache-busting proxy. *Computer communications*, 25(4), pp. 357–366.

Kim, K. (2001). Information seeking on the Web: effects of user and task variables. *Library and information science research*, 23(3), pp. 233–255.

King, N. S. (1991). Search characteristics and the effects of experience on end users of PaperChase. *College and research libraries*, 52(4), pp. 360–374.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association of Computing Machinery*, 46(5), pp. 604–632.

Koehler, W. C. (2002). Web page change and persistence: a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2), pp. 162–171.

Kraut, R. (1996). The Internet@home. *Communications of the ACM*, 39(12), pp. 32–35.

Kraut, R., Scherlis, W., Mukhopadhyay, T., Manning, J. & Keisler, S. (1996). The HomeNet field trial of residential Internet services. *Communications of the ACM*, 39(12), pp. 55–63.

Large, A. & Beheshti, J. (2000). The Web as a classroom resource: reactions from users. *Journal of the American Society for Information Science and Technology*, 51(12), pp. 1069–1080.

Large, A., Beheshti, J. & Rahman, T. (1999). Information seking on the Web: navigational skills og grade-six primary school students. In Woods, L., ed., *Proceedings of the 62nd ASIS annual meeting, Washington, DC, October 31–November 4, 1999*, vol. 36, pp. 84–97, American Society for Information Science, Medford, New Jersey: Information Today for ASIS.

Large, A., Beheshti, J. & Rahman, T. (2002). Gender differences in collaborative Web searching behavior: an elementary school study. *Information processing and management*, 38(3), pp. 427–443.

Lau, T. & Horvitz, E. (1999). Patterns of search: analyzing and modelling Web query refinement. In Kay, J., ed., *Proceedings of the seventh international conference on user modeling, Banff, Alberta, June 20–24, 1999*, pp. 119–128, User modeling inc., New York: Springer Wein, [also online], available from: <url:http://www.cs.usask.ca/UM99/Proc/lau.pdf>, [accessed 25 January 2001].

Lawrence, S. & Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280(5360), pp. 98–100, [3 April].

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the Web. *Nature*, 400(6740), pp. 107–109, [8 July].

Lazonder, A. W., Biemans, H. J. A. & Wopereis, I. G. J. H. (2000). Differences between novice and experienced users in searching information on the World Wide Web. *Journal of the American Society for Information Science*, 51(6), pp. 576–581.

Lesk, M. (1998). "Real World" searching panel at SIGIR 97. *SIGIR forum*, 32(1), pp. 1–4.

Li, J. H. (2000). Cyberporn: the controversy. *firstmonday*, 5(8), [online], available from: <url:http://firstmonday.org/issue5_8/li/index.html>, [accessed 1 Jul 2002].

Lin, S. & Belkin, N. J. (2000). Modeling multiple information seeking episodes. In Kraft, D. H., ed., *Proceedings of the 63rd ASIS annual meeting, Chicago, Illinois, November 12–16, 2000*, vol. 37, pp. 133–147, American Society for Information Science, Medford, New Jersey: Information Today for ASIS.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. London: MIT.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge University Press.

Marshall, I. & Roadknight, C. (1998). Linking cache performance to user behaviour. *Computer networks and ISDN systems*, 30(22–23), pp. 2123–2130.

Mathiesen, K. & Fallis, D. (2000). Book review: Steve Jones, ed. Doing Internet research: critical issues and methods for examining the net. *Journal of documentation*, 56(5), pp. 589–591.

Mayer, R. E. (1991). From novice to expert. In Helander, M., ed., *Handbook of human-computer interaction*, chap. 25, pp. 569–580, Amsterdam: Elsevier, second edition.

Mayo, E. (1933). *The human problems of an industrial civilization*. Cambridge, MA: Harvard University Press.

McKenzie, B. & Cockburn, A. (2001). An empirical analysis of Web page revisitation. In *Proceedings of the 34th Hawaii international conference on system sciences, Maui, Hawaii, January 3-6, 2001*, IEEE Computer Society, Los Alamitos: IEEE, [also online], available from: <url:http://www.cosc.canterbury.ac.nz/~andy/papers/hiccsWeb.pdf>, [accessed 11 June 2001].

Mead, S. E., Spaulding, V. A., Sit, R. A., Meyer, B. & Walker, N. (1997). Effects of age and training on World Wide Web navigation strategies. In *Proceedings of the Human Factors and Ergonomics Society 41st annual meeting, Albuquerque, New Mexico, September 22-26, 1997*, pp. 152–156, HFES, Santa Monica: Human Factors and Ergonomics Society.

Meadow, C. T. (2000). Letter: differences between novice and experienced users in searching information on the world wide web. *Journal of the American Society for Information Science*, 51(12), p. 1154.

Meyer, B., Sit, R. A., Spaulding, V. A., Mead, S. E. & Walker, N. (1997). Age group differences in World Wide Web navigation. In *CHI'97: proceedings of the conference on human factors in computing systems, Atlanta, Georgia, March 22-27, 1997*, pp. Electronic publications: late-breaking/short talks, ACM SIGCHI, New York: Association for Computing Machinery, [online], available from: <url:http://www.acm.org/sigchi/chi97/proceedings/short-talk/bm.htm>, [accessed 5 June 2001].

Mockapetris, P. (1987). RFC1034: Domain names: concepts and facilities. [Online], available from: <url:http://www.ietf.org/rfc/rfc1034.txt>, [accessed 28 November 2000].

Molloy, M. (1998). Letter: Searching the Web, continued. *Science*, 281(5374), pp. 176–177, [10 July].

Molyneux, R. E. & Williams, R. V. (2001). Measuring the Internet. In Williams, M. E., ed., *Annual review of information science and technology*, vol. 34, 1999/2000, chap. 2, pp. 287–339, Medford, NJ: Information Today for American Society for Information Science & Technology.

Moukdad, H. & Large, A. (2001). Users' perceptions of the Web as revealed by transaction log analysis. *Online information review*, 25(6), pp. 349–358.

Nesselroade, J. R. & Baltes, P. B., eds. (1979). *Longitudinal research in the study of behavior and development*. London: Academic.

Norman, D. A. (1983). Some observations on mental models. In Gentner, D. & Stevens, A. L., eds., *Mental models*, chap. 1, pp. 7–14, London: Lawrence Erlbaum.

Novak, T. P. & Hoffman, D. L. (1999). New metrics for new media: toward the development of of Web measurement standards. *World Wide Web journal*, 2(1), pp. 213–246, [also online], available from: <url:http://www2000.ogsm.vanderbilt.edu/novak/web.standards/webstand.html>, [accessed 22 June 2001].

Nowick, E. (2001). Using server logfiles to improve Website design. *Library philosophy and practice*, 4(1), [also online], available from: <url:http://www.uidaho.edu/ mbolin/nowick.pdf>, [accessed 1 November 2001].

Olson, M. A., Bostic, K. & Seltzer, M. (1999). BerkerleyDB. In *Proceedings of the USENIX annual technical conference, Monterey, California, June 6–11, 1999*, USENIX Association, [online], available from: <url:http://www.usenix.org/events/usenix99/full_papers/olson/olson.pdf>, [accessed 21 June 2001].

ONS (2002). Index of Internet connectivity. Report, Office for National Statistics, [online], available from: <url:http://www.statistics.gov.uk/statbase/>, [accessed 26 August 2002].

Penniman, W. D. & Dominick, W. D. (1980). Monitoring and evaluation of on-line information system usage. *Information processing and management*, 16(1), pp. 17–35.

Peters, T. A., Kurth, M., Flaherty, P., Sandore, B. & Kaske, N. K. (1993a). An introduction to the special section on transaction log analysis. *Library hi tech*, 11(2), pp. 38–40.

Peters, T. A., Kurth, M., Flaherty, P., Sandore, B. & Kaske, N. K. (1993b). Transaction log analysis. *Library hi tech*, 11(2), p. 37 et seq.

Pirolli, P. (2000). A Web site user model should at least model something about users. *Internetworking*, 3(1), [online], available from: <url:http://www.sandia.gov/itg/newsletter/mar00/critique_max.html>, [accessed 26 September 2001].

Pirolli, P. L. T. & Card, S. K. (1999). Information foraging. *Psychological review*, 106(4), pp. 643–675.

Pirolli, P. L. T. & Pitkow, J. E. (1999). Distributions of surfers' paths through the World Wide Web: empirical characterizations. *World Wide Web*, 2(1–2), pp. 29–45.

Pitkow, J. (1997). In search of reliable usage data on the WWW. *Computer networks and ISDN systems*, 29(8–13), pp. 1343–1355.

Pitkow, J. & Pirolli, P. (1999). Mining longest repeating subsequences to predict WWW surfing. In *Proceedings of the second USENIX symposium on Internet technologies and systems, Boulder, Colorado, October 11–14, 1999*, USENIX Association, [online], available from: <url:http://www.usenix.org/publications/library/proceedings/usits99/full_papers/pitkow/pitkow.ps>, [accessed 24 May 2001].

Pitkow, J. E. (1998). Summary of WWW characterizations. *Computer networks and ISDN systems*, 30(1–7), pp. 551–558.

Plewis, I. (1985). *Analysing change: measurement and explanation using longitudinal data*. Chichester: Wiley.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: automated library and information systems*, 14(3), pp. 130–137.

Qiu, L. (1993). Analytical searching vs. browsing in hypertext information retrieval systems. *Canadian journal of information and library science*, 18(4), pp. 1–13.

Rayward, W. B. (1996). The history and historiography of information science: some reflections. *Information processing and management*, 32(1), pp. 3–17.

Ross, N. C. M. & Wolfram, D. (2000). End user searching on the Internet: an analysis of term pair topics submitted to the Exite search engine. *Journal of the American Society for Information Science*, 51(10), pp. 949–958.

Salton, G. (1968). *Automatic information organization and retrieval*. London: McGraw-Hill.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. London: McGraw-Hill.

Schater, J., Chung, G. K. W. K. & Dorr, A. (1998). Children's internet searching on complex problems: performance and process analyses. *Journal of the American Society for Information Science*, 49(9), pp. 840–849.

Schechter, S., Krishnan, M. & Smith, M. D. (1998). Using path profiles to predict http requests. *Computer networks and ISDN systems*, 30(1–7), pp. 457–467.

Siegel, S. & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. London: McGraw-Hill, second edition.

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1998). Analysis of a very large AltaVista query log. SRC technical note 1998–014, Digital Systems Research Center, Palo Alto, [online], available from: <url:ftp://ftp.digital.com/pub/DEC/SRC/technical-notes/SRC-1998-014.ps>, [accessed 26 January 2001].

Silverstein, C., Marais, H., Henzinger, M. & Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR forum*, 33(1), pp. 6–12.

Smith, N. G. (1996). The UK national Web cache – the state of the art. *Computer networks and ISDN systems*, 28(7–11), pp. 1407–1414.

Spink, A. (1996). Multiple search sessions model of end-user behavior: an exploratory study. *Journal of the American Society for Information Science*, 47(8), pp. 603–609.

Spink, A., Griesdorf, H. & Bateman, J. (1999). A study of mediated successive searching during information seeking. *Journal of information science*, 25(6), pp. 477–487.

Spink, A., Jansen, B. J. & Ozmultu, H. C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet research: electronic networking applications and policy*, 10(4), pp. 317–328.

Spink, A., Jansen, B. J., Wolfram, D. & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3), pp. 107–109.

Spink, A., Wilson, T., Ellis, D. & Ford, N. (1998). Modeling users' successive searches in digital environments. *D-lib magazine*, 4(4), [online], available from: <url:http://www.dlib.org/dlib/april98/04spink.html>, [accessed 25 October 2000].

Spink, A., Wolfram, D., Jansen, B. J. & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), pp. 226–234.

Spink, A. & Xu, J. L. (2000). Selected results from a large study of Web searching. *Information research*, 6(1), [online], available from: <url:http://www.shef.ac.uk/~is/publications/infres/paper90.html>, [accessed 22 January 2001].

SPSS (1999). SPSS for Windows: release 10.0.5.

Su, Z., Yang, Q., Zhang, H., Xu, X. & Hu, Y. (2001). Correlation based document clustering using Web logs. In *Proceedings of the 34th Hawaii international conference on system sciences, Maui, Hawaii, January 3–6, 2001*, IEEE Computer Society, Los Alamitos: IEEE Computer Society, [online], available from: <url:http://dlib.computer.org/conferen/hicss/0981/pdf/09815019.pdf>, [accessed 11 June 2001].

Tabatabai, D. & Luconi, F. (1998). Expert-novice differences in searching the Web. In *Proceedings of the fourth Americas conference on information systems, Baltimore, Maryland, August 14–16, 1998*, pp. 390–392, Atlanta, Georgia: AIS, [also online], available from:

<url:http://www.isworld.org/ais.ac.98/proceedings/track07/tabatabai.pdf>, [accessed 23 May 2001].

Tashakkori, A. & Teddlie, C. (1998). *Mixed methodology: combining qualitative and quantitative approaches*. London: Sage.

Tauscher, L. & Greenberg, S. (1997a). How people revisit Web pages: empirical findings and implications for the design of history systems. *International journal of human-computer studies*, 47(1), pp. 97–137.

Tauscher, L. & Greenberg, S. (1997b). Revisitation patterns in World Wide Web navigation. In Pemberton, S., ed., *CHI'97: proceedings of the conference on human factors in computing systems, Atlanta, Georgia, March 22-27, 1997*, pp. 399–406, ACM SIGCHI, New York: Association for Computing Machinery.

Tauscher, L. M. (1996). *Evaluating history mechanisms: an empirical study of reuse patterns in World Wide Web navigation*. Master's dissertation, University of Calgary, [online], available from: <url:http://www.cpsc.ucalgary.ca/grouplab/papers/1996/96-Tauscher.Thesis/thesis.htm> [accessed 15 May 2001].

Thiébaut, D. (1989). On the fractal dimension of computer programs and its application to the prediction of the cache miss ratio. *IEEE transactions on computers*, 38(7), pp. 1012–1026.

Thomas, R. C. (1998). *Long term human-computer interaction: an exploratory perspective*. London: Springer.

Trybula, W. J. (1998). Data mining and knowledge discovery. In Williams, M. E., ed., *Annual review of information science and technology*, vol. 32, 1997, chap. 4, pp. 197–230, Oxford: Learned Information for American Society for Information Science.

UCLA Center for Communication Policy (2002). The UCLA Internet report. [Online], available from: <url:http://ccp.ucla.edu/pages/internet-report.asp>, [accessed 5 July 2002].

Voldman, J., Mandelbrot, B., Hoevel, L. W., Knight, J. & Rosenfeld, P. (1983). Fractal nature of software–cache interaction. *IBM journal of research and development*, 27(2), pp. 164–170.

W3C Web characterization activity (1999). Concept of "user". Mailing list discussion, World Wide Web Consortium, [online], available from: <url:http://www.w3.org/WCA/>, [accessed 28 August 2002].

Wall, L., Christiansen, T. & Schwartz, R. L. (1996). *Programming Perl*. Cambridge, MA: O'Reilly, second edition.

Wang, P. (2001). Methodologoies and methods for user behavioral research. In Williams, M. E., ed., *Annual review of information science and technology*, vol. 34, 1999/2000, chap. 2, pp. 53–99, Medford, NJ: Information Today for American Society for Information Science & Technology.

Wersig, G. (1993). Information science: the study of postmodern knowledge usage. *Information processing and management*, 29(2), pp. 229–239.

Williams, T. & Kelley, C. (1998). *gnuplot: an interactive plotting program*. Version 3.7, manual prepared by D. Crawford.

Wilson, T., Ellis, D., Ford, N. & Foster, A. (2000). Uncertainty in information seeking. Research report 59, Library and Information Commission, Boston Spa.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of documentation*, 55(3), pp. 249–270.

Wolfram, D., Spink, A., Jansen, B. J. & Saracevic, T. (2001). Vox populi: the public searching of the Web. *Journal of the American Society for Information Science and Technology*, 52(12), pp. 1073–1074.

Yuan, W. (1997). End-user searching behavior in information retrieval: a longitudinal study. *Journal of the American Society for Information Science*, 48(3), pp. 218–234.

Zhang, J. & Korfhage, R. R. (1999). A distance and angle similarity measure method. *Journal of the American Society for Information Science*, 50(9), pp. 772–778.

Zipf, G. (1972). *Human behaviour and the principle of least effort: an introduction to human ecology*. New York: Hafner, facsimile of 1949 edition.

Glossary

AltaVista-Excite sample The subset of the Web log which includes all the search-queries submitted to the AltaVista and Excite 'search-engines' together with the search-parts submitted.

average query-click proportion The user-characterization given by the ratio of the student-user's total number of query-clicks to the total number of clicks.

average search-query count The user-characterization given by the ratio of the student-user's total number of search-queries to the total number of search-sessions.

average search-query proportion The user-characterization given by the ratio of the student-user's number of search-query-clicks to the number of clicks.

average search-session proportion The 'search-engine' usage characterization given by the ratio of the number of search-sessions to the number of sessions.

average search-term count The user-characterization given by the ratio of the student-user's total number of search-terms to the total number of search-queries.

average session click rate The user-characterization given by the ratio of the student-user's total number of clicks to the number of sessions.

average session-conformance The user-characterization given by the ratio of the student-user's total session-conformance to the number of sessions.

average Webhost-persistence The user-characterization given by the ratio of the student-user's sum total of session Website-repertoire to sum total of session Webhost-repertoire.

average Website-re-request rate The user-characterization given by the ratio of the student-user's total number of clicks to the sum total of session Website-repertoire.

by-click see Web log analysis.

by-session see Web log analysis.

by-user see Web log analysis.

cache-busting Techniques employed by Web-servers to overcome the effects of caching in particular to force the Web-client to obtain a fresh copy of the Webpage from the Web-server on each occasion that it is requested.

caching The temporary storage either by a local client cache or in a network cache of files obtained from Web-servers in order to provide a more efficient response to a Web-client request.

click A request to a Web-server for a file (or files) identified by a url.

click rate The number of clicks during some period.

clickstream see trace.

client see client-server.

client-server The client-server computing model partitions functionality and workload. In respect of the Internet the client computer program, for example a Web client or browser, is operated by the user. Each Web browser communicates with a Web server which is able to respond to requests and sends files to the client. Client-side and server-side refers to one or other perspective of the complete client-server system.

client-side see client-server.

collective-popularity see popularity.

conditional analysis The name used here to describe a longitudinal-developmental analysis using conditional regression to investigate the phenomenon of change. Conditional analysis models the value of a user's characteristic at a later time as a function of the value of the characteristic at an earlier time so that in the absence of change the function is an identity.

conditional-conformance The compound conformance user-attribute (either conformant or eclectic) describing the change in conformance.

conditional-regression see conditional analysis.

conditioned url-string see conditioning.

conditioning The string and character manipulation procedure (munging) used to standardise the url-string representation of the urls submitted to Web-servers in order to compare these more reliably. In particular host-names are munged.

conformance The user-attribute which describes whether a student-user always visits at least one of the more popular Websites during every session (conformant) or not (eclectic).

conformant see conformance.

consent see ethical practice.

consistent Findings or conclusions are described as being consistent if they apply to each of the study-years and to each of the user-attribute partitions.

cookie A coded text file generated by a Web-server and stored by the Web-client. Cookies can be used to track Web usage.

Data Protection Act Principle based UK legislation to regulate obtaining, storing, using and passing on personal data.

domain name server An Internet service based on a hierarchy of cooperating name servers which regulates the host-names of servers and in particular provides a translation of host-name to Internet Protocol address. The official name is a unique name for the each server within the Internet although each server may also be known by many alias names.

DNS see domain name server.

double-conformant see conditional-conformance.

double-eclectic see conditional-conformance.

eclectic see conformance.

embedded Files such as images which are referenced within the hypertext mark-up of a Webpage and are therefore requested implicitly by the Web-client.

energetic Student-users are described as being more energetic when they make more clicks.

epoch The time reference scheme used by Unix and Unix-like systems based on counting seconds elapsed from 00:00:00 1 January 1970.

ethical practice The consent model is applied in respect of small scale Web research but larger scale Web research based on Web log data provided by a third party uses the permission model, see Chapter two.

expire flag see global history.

extensible markup language A replacement for hypertext markup language which provides more flexibility for authors.

gender A user-attribute.

ghar see global history archive file.

global history The history mechanism used by the Netscape Navigator browser, for example, to mark hyperlinks as having been visited by maintaining a visit count of visits to each url. The expire flag allows these links to be restored to unvisited when the visit count is reset to zero.

global history archive file This is a daily compressed archive file from the institution which contains global history files.

HCI see human-computer interaction.

history mechanisms see global history.

host-munging see conditioning.

host-name The canonical representation of Web-server within a url, for example in <url:http://host.com/somefile.html> the host-name is 'host.com'.

html see hypertext markup language.

human-computer interaction The study of (usually low level) interaction phenomena between users and computer systems.

hypertext markup language A method used to describe Webpages and in particular to embed clickable instructions or hyperlinks to other Webpages.

individual-popularity see popularity.

information retrieval Techniques for identifying relevant documents held in a storage system in response to a user's search-query. Each search-query comprises search-terms and is specially formulated according to the protocol of the system. The information retrieval process is called here the search-query paradigm.

internal-anchor An html technique which allows a hyperlink to refer to another location within the same html document. Internal-anchors trail url-strings and are demarcated by '#'.

Internet A network of distributed but connected computers which cooperate and share services and files by adhering to common standards and protocols.

Internet Protocol The address like identification scheme which facilitates routing messages within the Internet from computer to computer.

IP see Internet Protocol.

IR see information retrieval.

joint-session-rate The sum of the session rate during each study-year. Student-users are partitioned into smaller and larger based on comparing their joint-session-rate with the mean value of joint-session-rate.

larger An attribute partition for session rate and joint-session-rate. For example larger student-users have greater than average session rates.

link-clicking A user following hypertext links (not containing search-parts) as opposed to query-clicking.

longitudinal-developmental Longitudinal studies which focus on how an individual changes over time.

no-change line The identity function, see conditional-analysis.

novice A student-user who is relatively lacking in experience. The notion of novice does not entail any knowledge of a student-user's expertise.

novice-effect A user-characteristic is said to exhibit the novice-effect when change of the characteristic for novice student-users is an exaggeration of the change for more experienced student-users. Since this is equivalent to the rate of change reducing as users become more experienced then the novice-effect can be likened to a learning curve in respect of the user-characteristic.

official name see domain name server.

online public access catalogue A library management system which allows library users to interrogate the library catalogue from computer terminals.

OPAC see online public access catalogue.

path lengths The number of successive requests for Websites by a user. Depending on context these may be different Websites. The criterion for starting a new path may also vary.

permission see ethical practice.

personal data see Data Protection Act.

popularity As in Website popularity, the number of different users who visit the Website during some period. Individual-popularity is the user frequency while relative-individual-popularity is the relative frequency.

query In an information retrieval sense, a query is a search request to an IR system. In a Web log sense, a query relates to any url which contains a search-part.

query-click A click or Web request where the url contains a search-part.

query-click proportion The ratio of the query-click rate to the click rate.

query-click rate The number of query-clicks during some period.

query-clicking A user submitting queries as opposed to link-clicking.

query-session A session which contains a query-click.

query-session proportion The ratio of query-sessions to sessions.

rare Websites that are not among the top one thousand most visited Websites.

real world As opposed to experimental. Investigations of users' information behaviour where this is naturalistic and relates to the users' own information needs.

relative-individual-popularity see popularity.

repertoire The cardinality of the vocabulary set.

search-engine The popular name for a Web information location service.

search-part The terminating part of a url-string which is demarcated by a '?' character.

search-query As a contraction of Web search-query this is an information retrieval type query used in the context of requesting information from one of the fourteen principal Web information location service identified within the study.

search-query count The number of search-query-clicks submitted during a session.

search-query-click A query-click submitted by a student-user which is to one of the fourteen principal Web information location service identified within the study.

search-query paradigm see information retrieval.

search-session A session which includes a search-query.

search-session proportion The user-characterization given by the ratio of the student-user's total number of search-sessions to the number of sessions.

search-session rate The count of the number of search-sessions during a study-year.

search-term count The number of search-terms in a search-query.

search-term The constituents of the search-query. Since these are usually delimited by ' ' (a space) then each search-term can be thought of as a word.

search-user A student-user who has a search-session.

server see client-server.

server-side see client-server.

session Generally a time delimited sample of information behaviour. Here each session comprises the Web log transactions for a student-user during a single day.

session click rate The number of clicks during a session.

session query-click rate The number of query-clicks during a session.

session rate The number of sessions during a study-year.

session vocabulary The set of different Websites or Webhosts visited during a session.

session-by-session see Web log analysis.

session-conformance A metric computed for each session using the vector model which compares the session with the typical session.

session-rate A user-attribute; student-users are partitioned into smaller and larger based on comparing their session-rate with the mean value of session-rate.

session-to-session see Web log analysis.

singleton Singleton search-queries are search-queries which consist of a single search-term.

smaller see session-rate

structural It is hypothesised that the longitudinal investigation of how student-users use the Web will reveal instances of change in the Web's information seeking affordances. This change is called structural in order to distinguish it from any individual change by student-users.

student-users All of the students at the institution are potentially included in the Web log. Those 1,050 students from the two cohorts who used the Web on at least two occasions during each study-year and are therefore included in the investigation are referred to student-users in order to distinguish them,

study-year The first and second academic years of the two year long period of the investigation.

study-year vocabulary The union of the session vocabulary over all the sessions in a study-year.

term Terms (or words) are used to construct search-queries.

term frequency*inverse document frequency A generic approach to matching documents to queries in the vector model is to give more attention to documents where the term appears more often (the term frequency) but to give less attention to terms which occur in many documents within the collection (the document frequency).

tf*idf see term frequency*inverse document frequency

timestamp A general procedure within transaction logging systems to record the time of occurrence of an event as well as a synopsis of the event.

TLA see transaction log analysis.

trace Within the computer science and electrical engineering communities, the log record of the clickstream of successive Web requests is sometimes called a trace.

trace analyses As Web log analysis but in respect of the clickstream trace.

trajectory As used here, a student-user's trajectory is the relationship between repertoire and cumulative click rate. The gradient of the fitted trajectory function therefore indicates the repertoire growth characteristic. This is discussed in Chapter three.

trajectory analysis This is the (statistical) analysis of student-users' trajectories, in particular of the gradient of the fitted trajectory functions. Trajectory analysis can be applied to either Websites or Webhosts.

transaction log analysis Within library and information science transaction log analysis refers to an analysis of the bibliographic searching using OPACs. Matters of interest include the comparative usage of system features.

uniform resource locator Part of the Internet procedure used to request files and services. Each url includes the host-name of the server from which the file or service is requested so that the request can be routed to the required server.

url see uniform resource locator.

url-string A particular (non-unique) character-string representation of the url.

user-attributes Cohort, gender and session-rate attributes are used to partition the student-users so that, for example, how men student-users locate Web information can be compared with how women student-user locate Web information.

user-characterization Web information seeking user-characterization metrics are used to describe and compare aspects of how individual student-users locate Web information.

user-id Each student-user has a unique user-id which is used in combination with personal password to gain initial access to the institution's network. A global history file is maintained for each student-user and is stored in that user's 'private' file space.

vector model An information retrieval model which decomposes documents into their constituent terms and represents each document in a collection as a term vector having components equal to the term frequency within the document.

visit count The global history mechanism maintains a cumulative count of requests to each Website whether or not this is satisfied from the local browser cache. The count is called the visit count.

vocabulary Originally the set of different words in a text, more generally the set of different terms in a collection. The Website vocabulary, for example, is a set of different Websites.

Web The common contraction of 'World Wide Web', the name selected by Tim Berners-Lee to describe the novel Internet service which he was proposing.

Web client see client-server.

Web information This phrase is used here in a restrictive sense to exclude a range of services such as email which can be accessed by means of a Webpage 'front end' but which are not essentially 'Web'.

Web information location services This description is used to apply to all of the popular Web 'search-engines' including those which do not use robots to crawl the Web.

Web log The Web log is derived by decumulating and cleaning (by conditioning) collections of successive global history files from 1,050 student-users.

Web log analysis Analysis of the Web transactions recorded in the Web log in order to reveal information about both how the Web is used and how users use the Web. This can be undertaken from different perspectives depending on the purpose of the analysis. For example the Web log can be analysed by-click, by-session and by-user. As well as micro analyses which provide session-by-session information, the Web log can be analysed in a macro sense, session-to-session, so that successive session phenomena are revealed.

Web search-queries see search-query.

Web server see client-server.

Webhost The conditioned host-name.

Webhost-persistence see user-characterization

Webhost-trajectory see trajectory analysis.

Webpage A Webpage is an assemblage of one or more files rendered by the Web browser as a single viewable document which is possibly scrollable in response to the user making a request for Web information to a Website.

Website A Website is generally the Web resource to which a request for Web information results in the display of a Webpage by the Web browser. The Website is derived by conditioning the url-string and excludes search-parts and internal-anchors. When the url includes neither of these then the Website is equivalent to the url-string by which the Webpage is requested but with the host-name replaced by the unique official name.

Website-re-request rate This metric is the ratio of the number of clicks to the Website-repertoire.

Website-repertoire The cardinality of the Website-vocabulary (set of different Websites).

Website-trajectory see trajectory analysis.

xml see extensible markup language.

Appendices

Appendix A

How do student-users locate Web information?

Institutional background

Tables A.1, A.2 and A.3 show extracts from Tables 8a, 8g and 9a of *Students in higher education institutions* (Higher Education Statistics Agency, 1999, 2000, 2001). These describe the subject areas studied by the full-time undergraduates attending the institution during the study-years and the gender and age distribution information which is available in respect of the student cohorts from which the student-users are drawn. Students of unknown age are included in the totals but not in the breakdowns.

The institutional background to the study is discussed to in Chapter one.

Subject area	1998/1999	1999/2000
Subjects allied to medicine	115	110
Biological sciences	538	610
Agriculture & related subjects	8	-
Physical sciences	247	240
Computer science	533	500
Architecture, building & planning	101	70
Social economics & political studies	216	250
Business & administrative studies	1,288	1,340
Librarianship & information science	173	190
Languages	216	230
Humanities	177	170
Creative arts & design	664	690
Education	456	410
Combined	649	750
Total full-time undergraduates	5,381	5,570

Table A.1: The institution's full-time undergraduates by subject area

The institution	1997/1998	1998/1999
All higher education students	7,971	7,899
Total undergraduates	6,807	6,999
men	2,885 or 42%	2,981 or 43%
women	3,922 or 58%	4,018 or 57%
Total first year full-time undergraduates	2,103	2,345
aged under 21	1,434 or 68%	1,718 or 73%
age 21 or older	668 or 32%	625 or 27%

Table A.2: The gender and ages of students at the institution

The UK	1997/1998	1998/1999
Total undergraduates	1,413,063	1,442,417
men	46.8%	46.1%
women	53.1%	53.9%
Total first year full-time undergraduates	394,207	392,498
aged under 21	269,540 or 69%	271,440 or 69%
age 21 or older	123,478 or 31%	120,305 or 31%

Table A.3: The gender and ages of UK undergraduates

Appendix B

A method for discovering how student-users locate Web information

Session frequencies

Tables B.1 and B.2 show the 21,366 and 25,192 daily Web information seeking sessions during study-years one and two respectively cross-tabulated according to the cohort attribute of the originating student-user and the click-rate attribute. The click-rate attribute is *smaller/larger* depending on whether or not is $<$ or $\geqslant$ the mean session click rate during the study-year.

The classification attributes 'cohort' and 'click rate' are not independent. This is discussed in Chapter three.

Web vocabulary and trajectory analysis

Figures B.1 and B.2, and Figures B.3 and B.4, illustrate for study-years one and two respectively overlays of the Website and Webhost trajectories of student-users.

Trajectory analysis is discussed in Chapter three.

Sessions

Student-user cohort	Study-year one click rate		Both
	smaller	larger	
1997 cohort	5,730 sessions	2,893 sessions	8,623 sessions
1998 cohort	9,664 sessions	3,079 sessions	12,743 sessions
Study-year one	15,394 sessions	5,972 sessions	21,366 sessions

Table B.1: Cross-tabulation of session frequency by cohort and click rate during study-year one

Student-user cohort	Study-year two click rate		Both
	smaller	larger	
1997 cohort	6,521 sessions	3,314 sessions	9,835 sessions
1998 cohort	10,823 sessions	4,534 sessions	15,357 sessions
Study-year two	17,344 sessions	7,848 sessions	25,192 sessions

Table B.2: Cross-tabulation of session frequency by cohort and click rate during study-year two

Trajectories

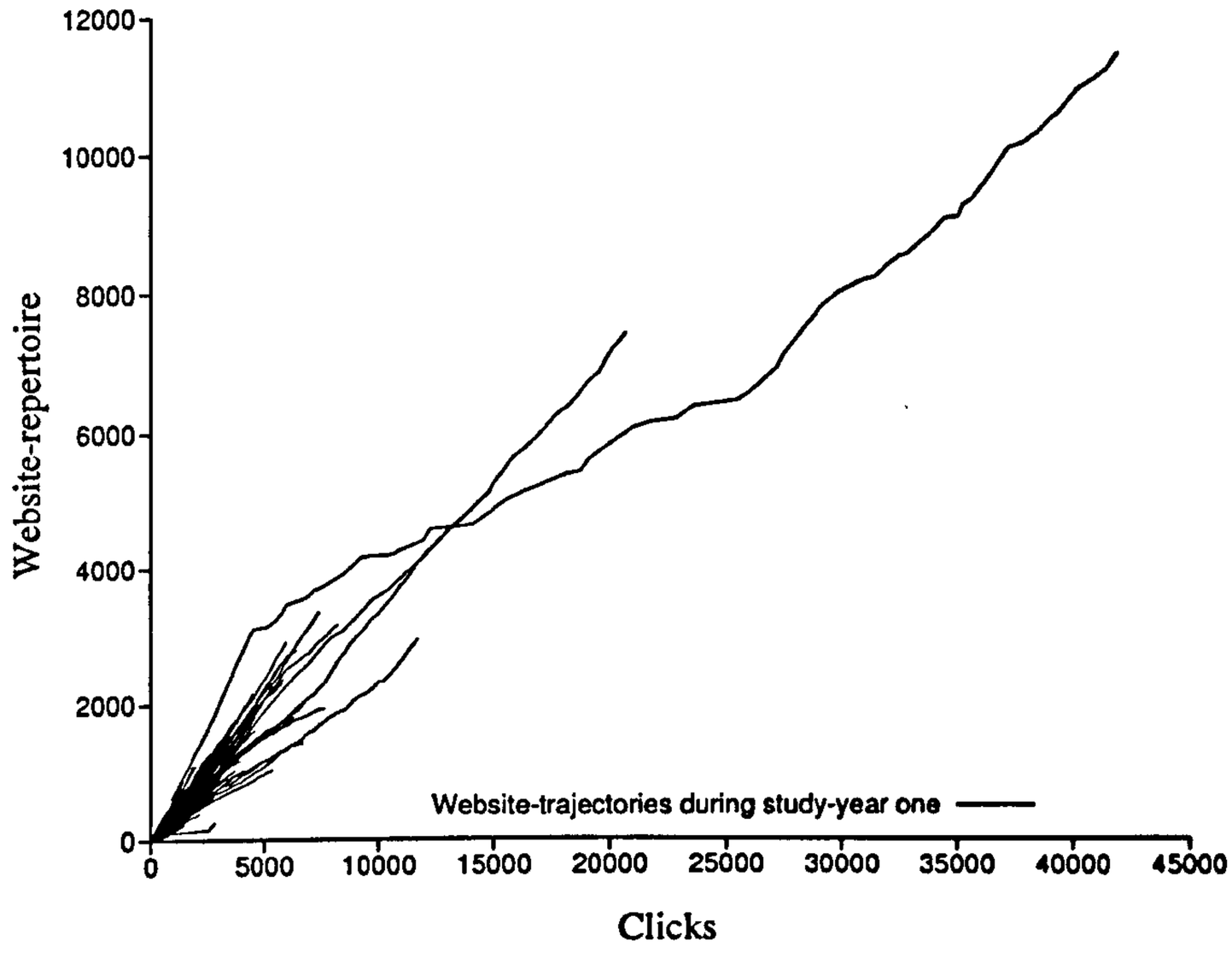

Figure B.1: Overlay of Website-trajectories during study-year one

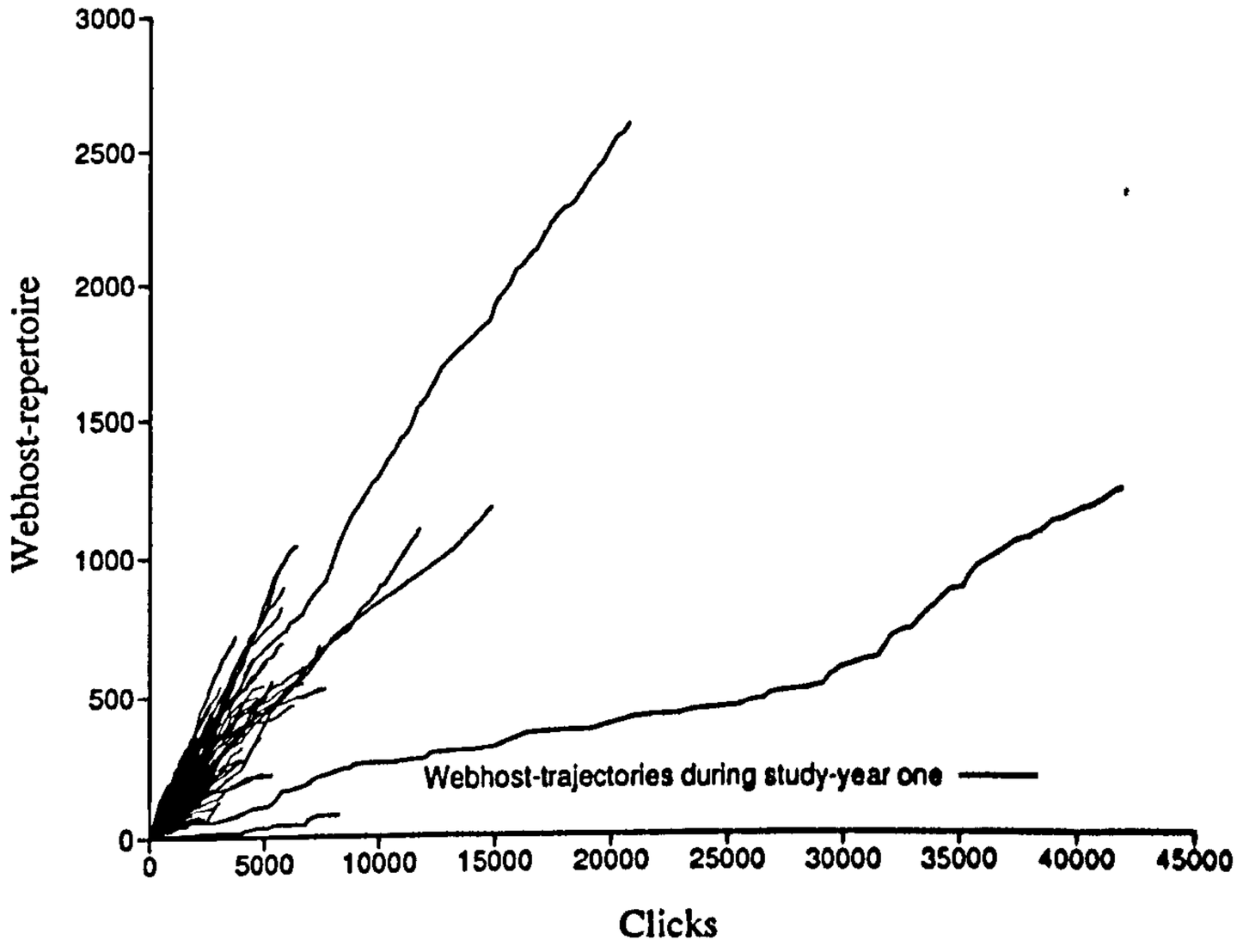

Figure B.2: Overlay of Webhost-trajectories during study-year one

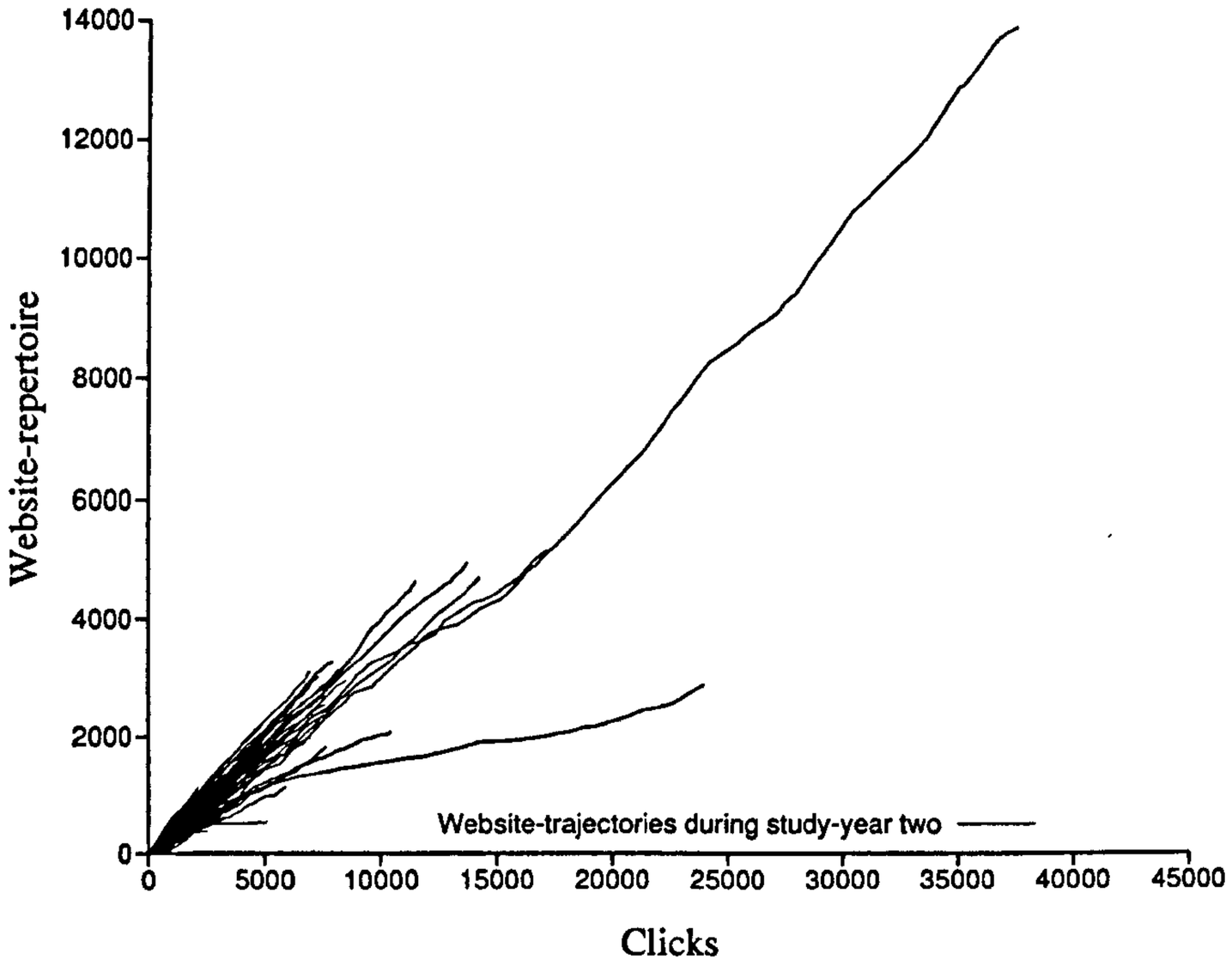

Figure B.3: Overlay of Website-trajectories during study-year two

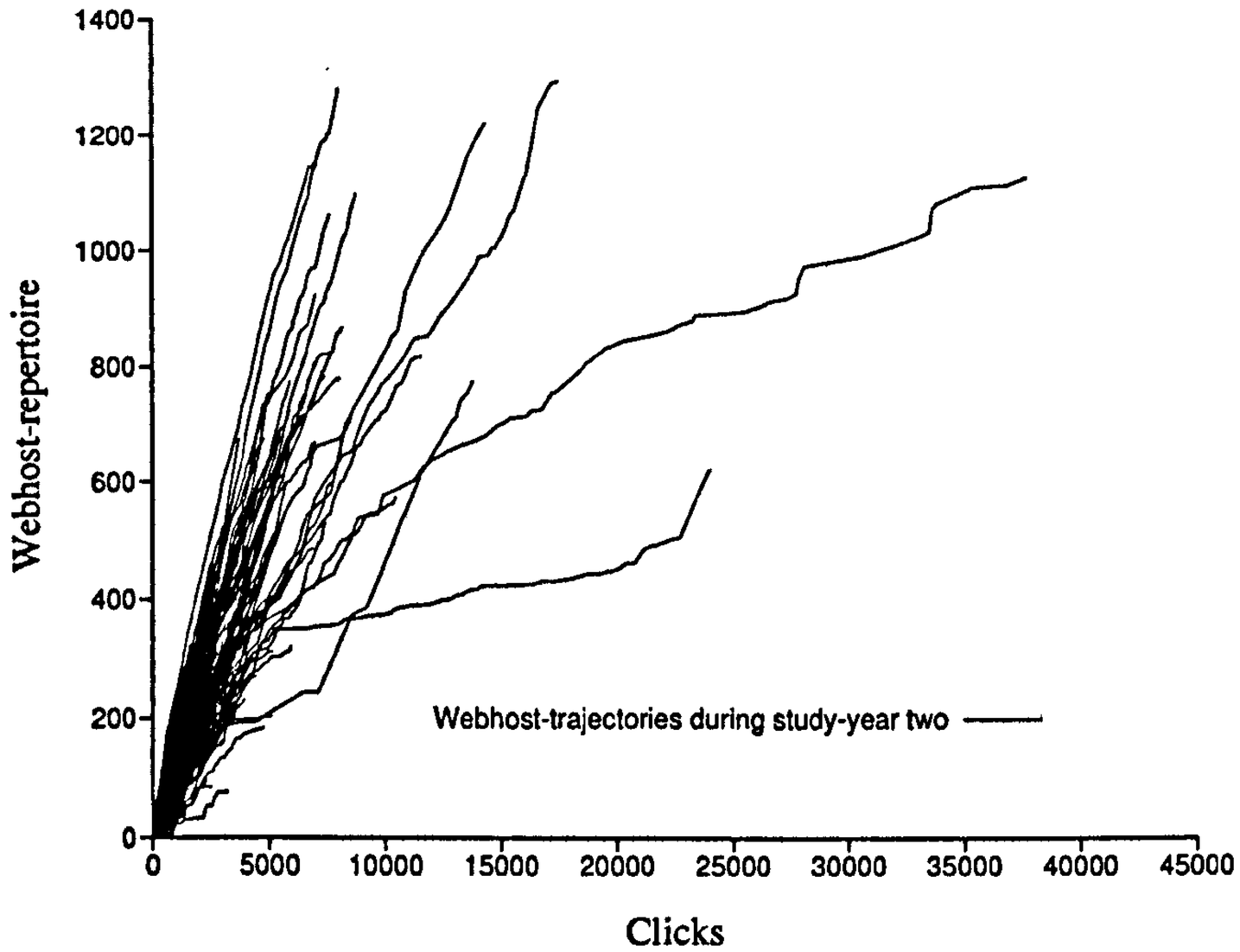

Figure B.4: Overlay of Webhost-trajectories during study-year two

Appendix C

How do student-users use the Web?

User-characterization

Figure C.1 illustrates the frequency distributions of query-session rates and Tables C.1 and C.2 report the query-click, search-query-click, query-session and search-session Web log frequencies.

Figures C.2 and C.3 illustrate the scattergrams of student-user's average session click rate and average Website-re-request rate discussed in the context of average Website-re-request rates in Chapter four.

Figures C.4 and C.5 illustrate the repertoire frequency distributions discussed in Chapter four. Figures C.6 to C.11 illustrate the scattergrams of student-user's average session click rate, average Website-re-request rate, and repertoire discussed in Chapter four in relation to average Webhost-persistence.

Figures C.12 and C.13 illustrate scattergrams of average session-conformance and session-conformance range.

Similarities and differences

Tables C.3 to C.5 relate to the similarities and differences between student-users in different user-attribute (gender, session-rate and conformance) partitions which are discussed in Chapter four.

Website popularity

The Zipf distributions of individual-popularity are illustrated in Figures C.14 and C.15 and the 'top-twenty' Websites are shown in Tables C.6 and C.7.

Figures C.16 and C.17 illustrate collective-popularity. Website popularity is discussed in Chapter four.

Average query-click proportion

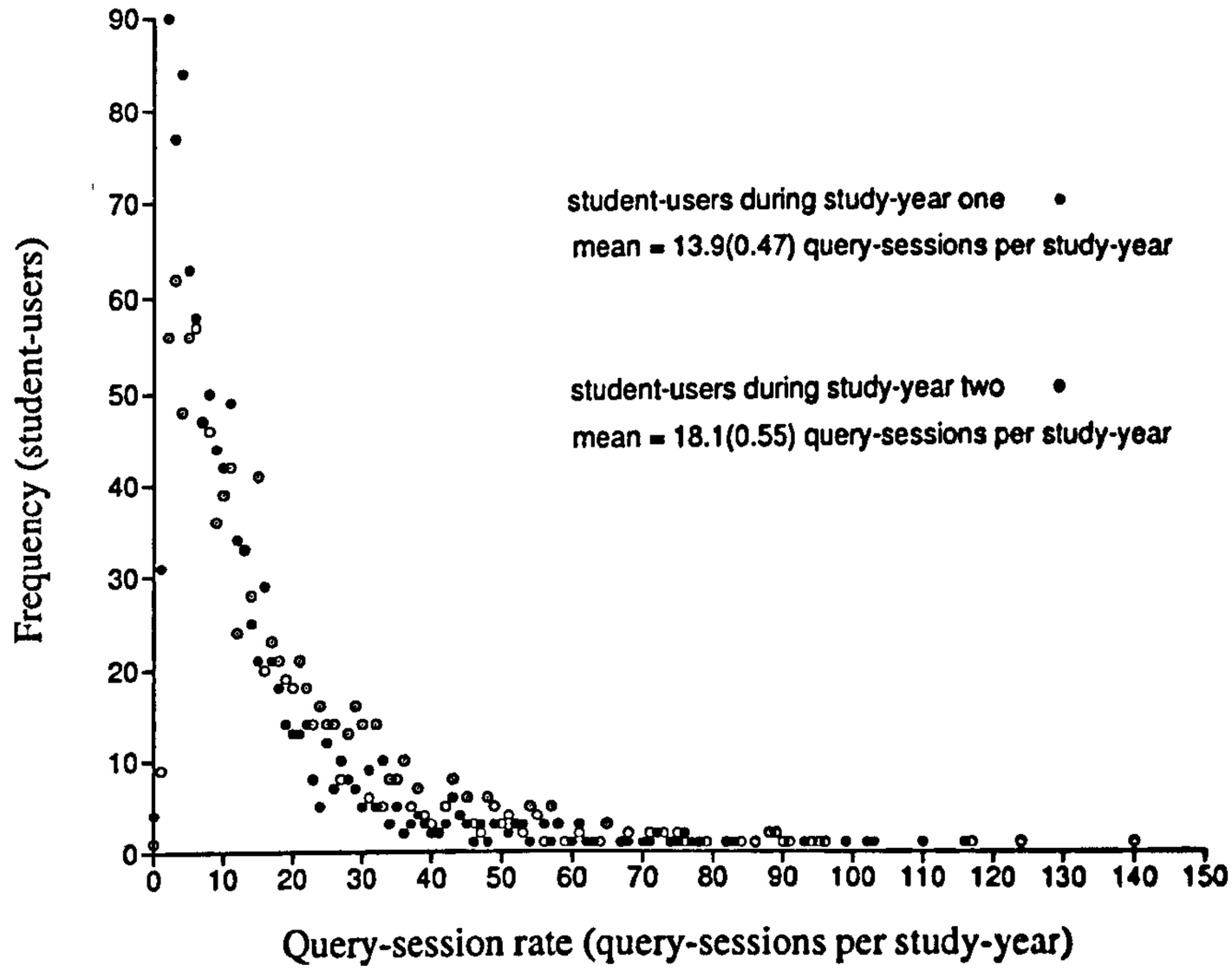

Figure C.1: Freqency distributions of student-user's query-session rate

Study-year	Attribute		
	clicks	query-click	search-query-click
study-year one	758,636 clicks	194,232 query-clicks	86,853 search-query-clicks
study-year two	1,231,852 clicks	389,136 query-clicks	151,761 search-query-clicks
Overall	1,990,488 clicks	583,368 query-clicks	238,614 search-query-clicks

Table C.1: Cross-tabulation of Web log click attribute frequencies by study-year and attribute

Study-year	Attribute		
	sessions	query-sessions	search-sessions
study-year one	21,366 sessions	14,551 query-sessions	10,483 search-sessions
study-year two	25,192 sessions	18,996 query-sessions	12,311 search-sessions
Overall	46,558 sessions	33,547 query-sessions	22,794 search-sessions

Table C.2: Cross-tabulation of Web log session attribute frequencies by study-year and attribute

Average Website-re-request rate

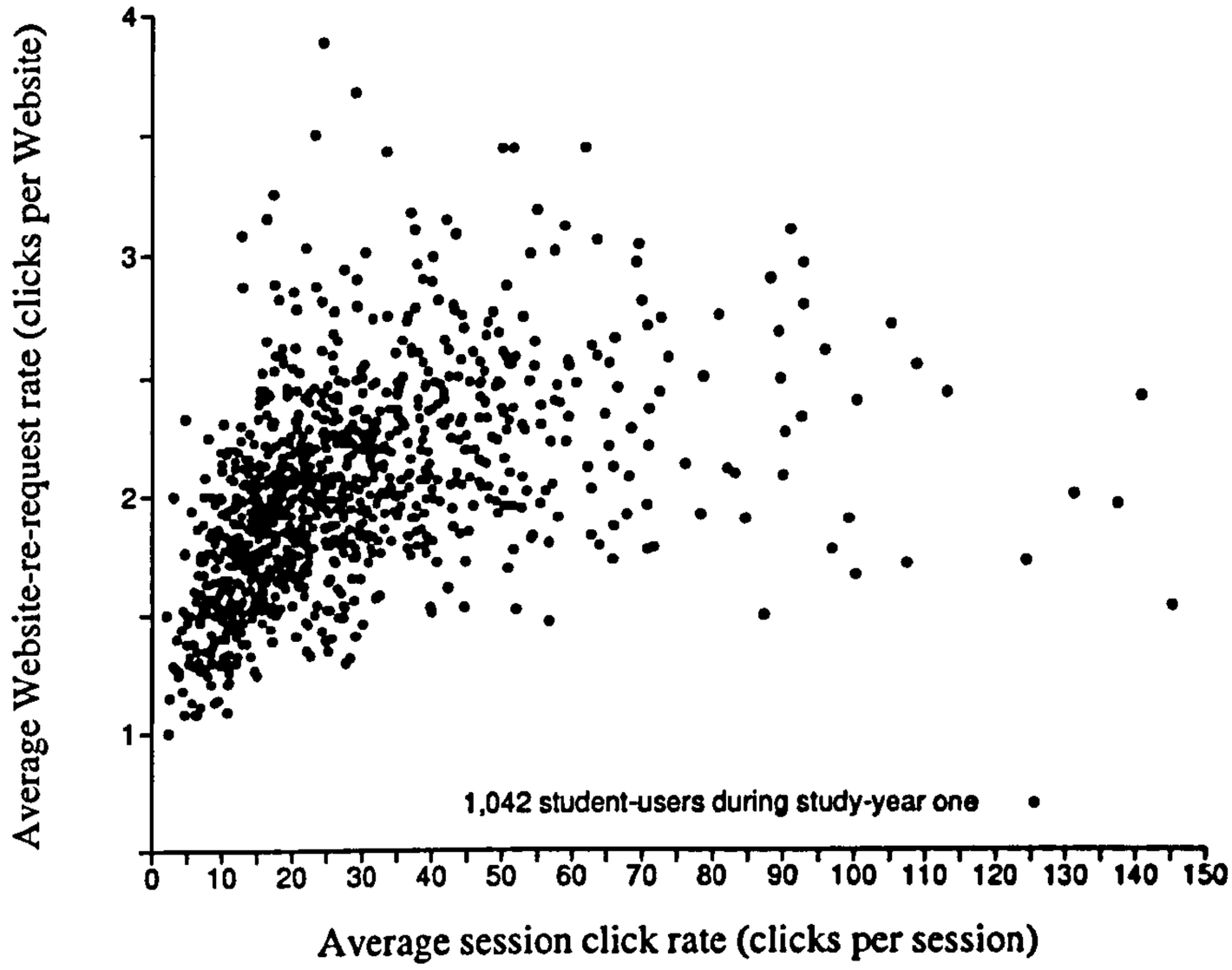

Figure C.2: Scattergram of 1,042 student-user's average session click rate and average Website-re-request rate during study-year one (range illustrated up to 150 clicks per session and four clicks per Website)

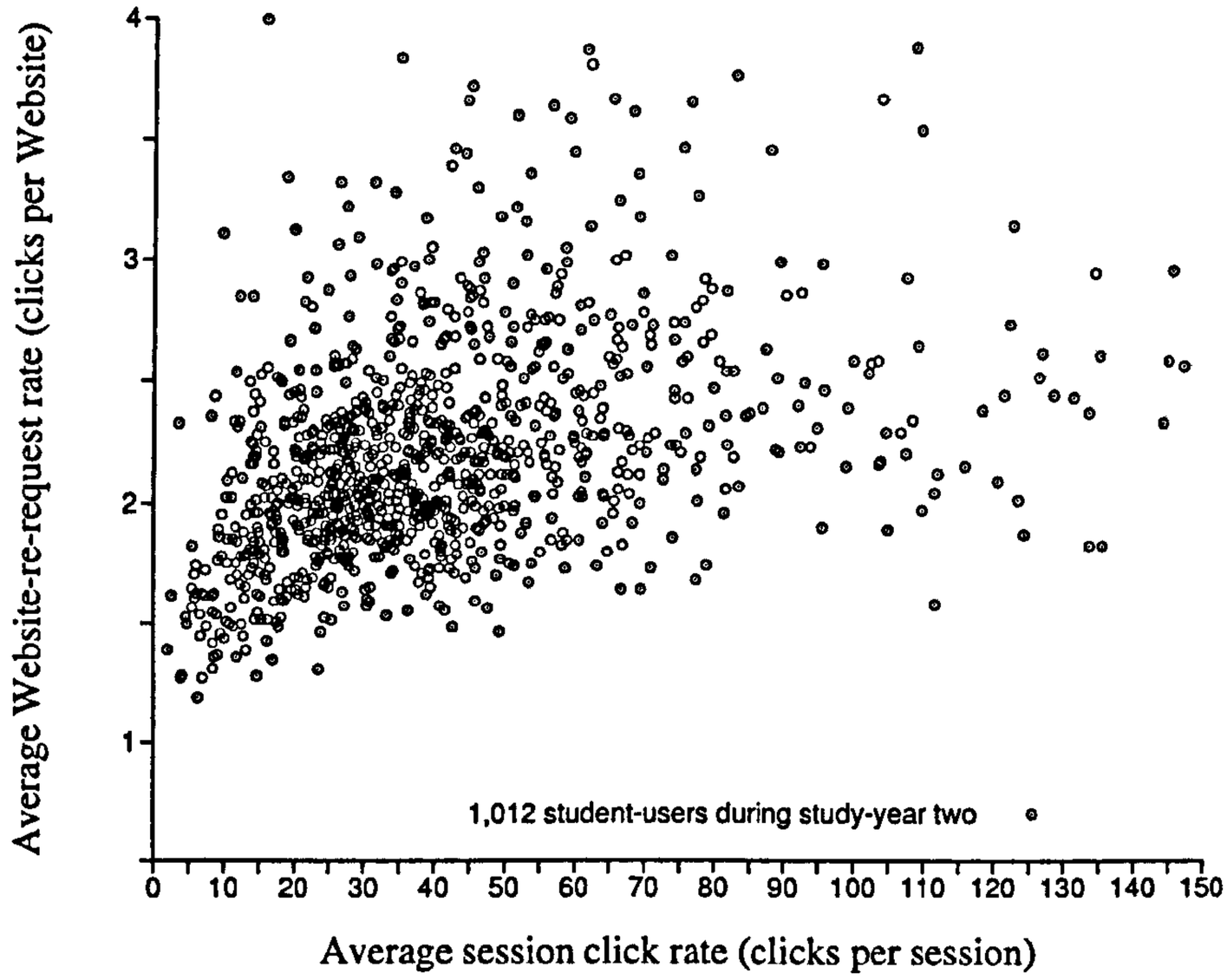

Figure C.3: Scattergram of 1,012 student-user's average session click rate and average Website-re-request rate during study-year two (range illustrated up to 150 clicks per session and four clicks per Website)

Average Webhost-persistence

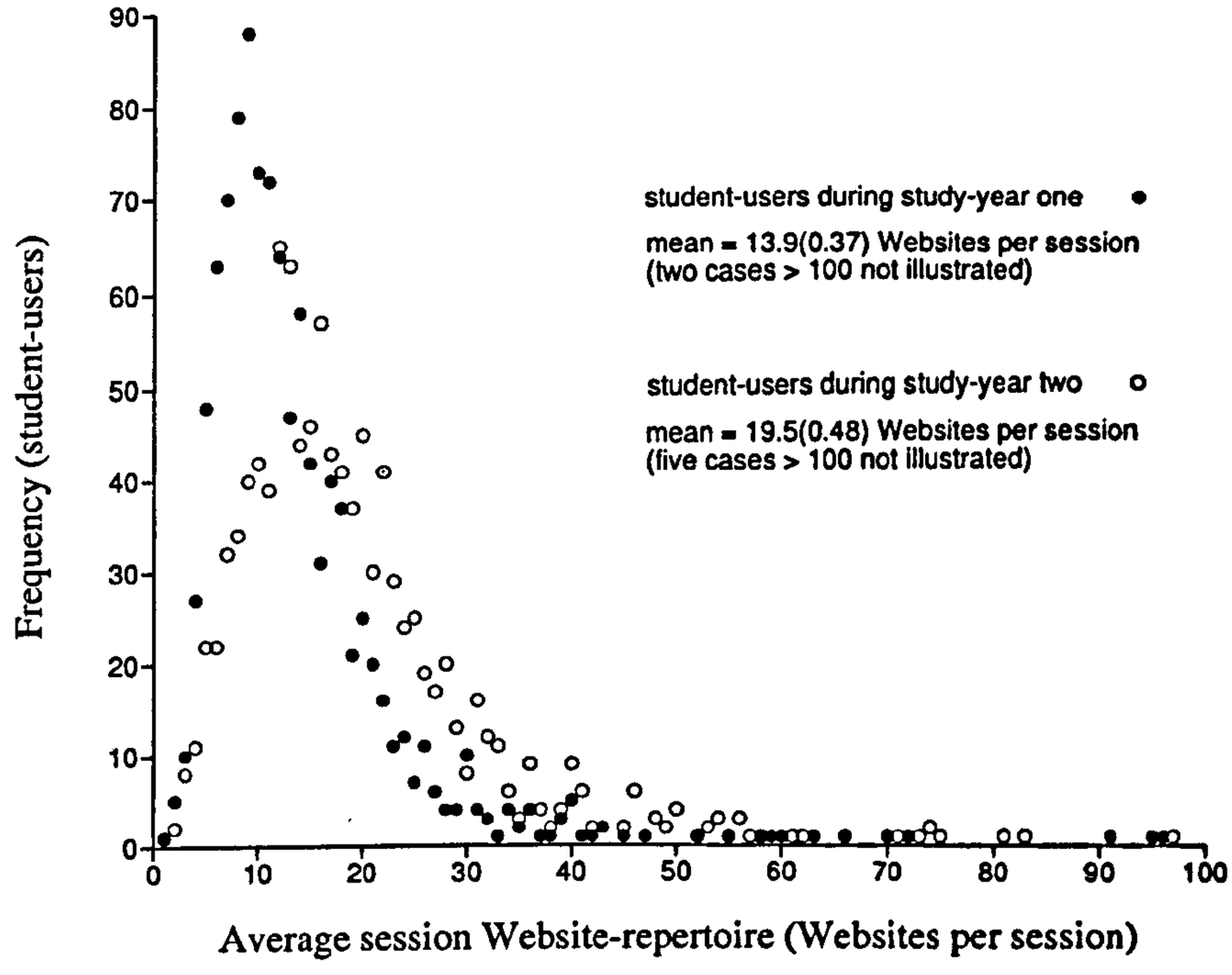

Figure C.4: Frequency distributions of student-user's average session Website-repertoire (range illustrated up to 100 Websites per session)

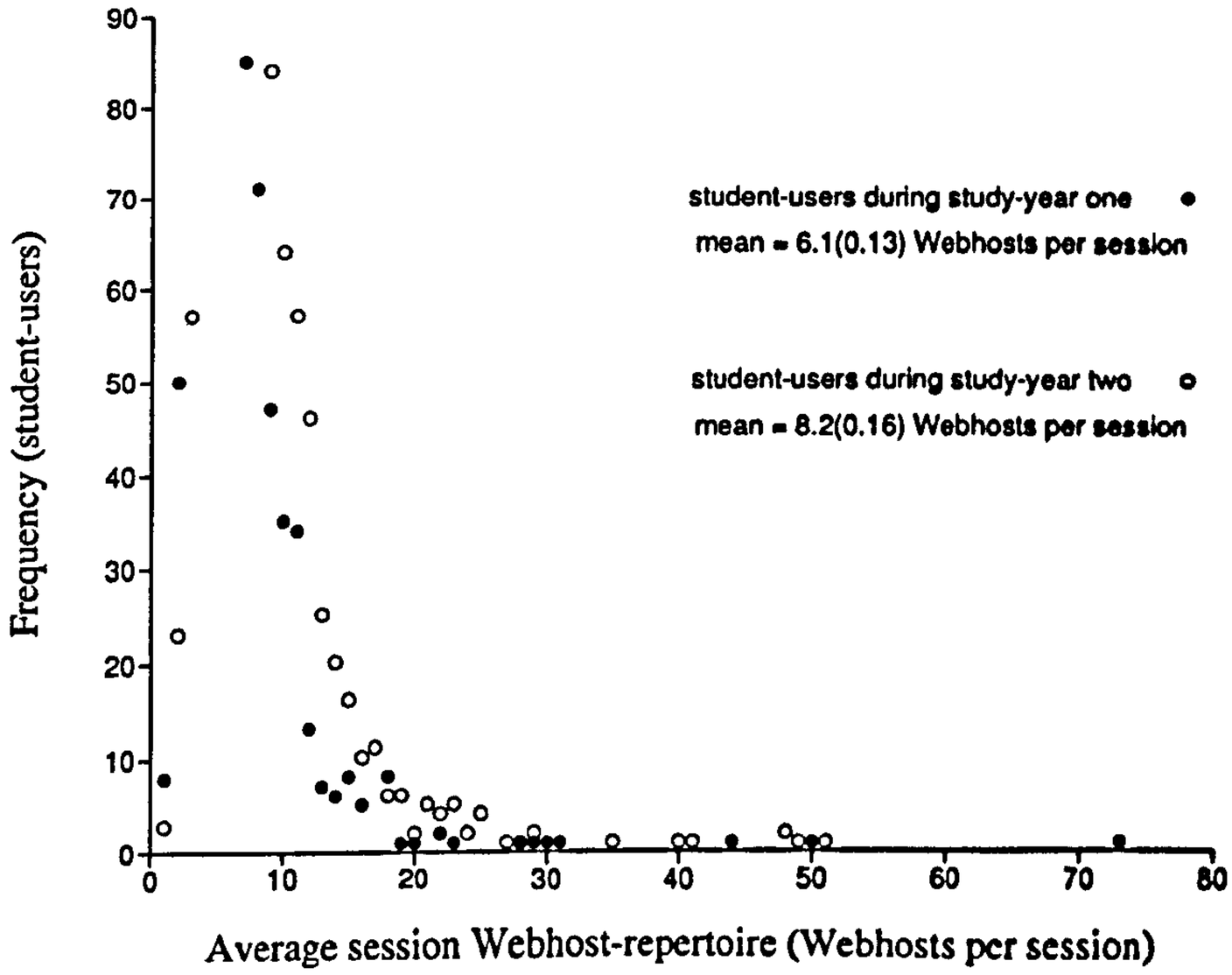

Figure C.5: Frequency distributions of student-user's average session Webhost-repertoire

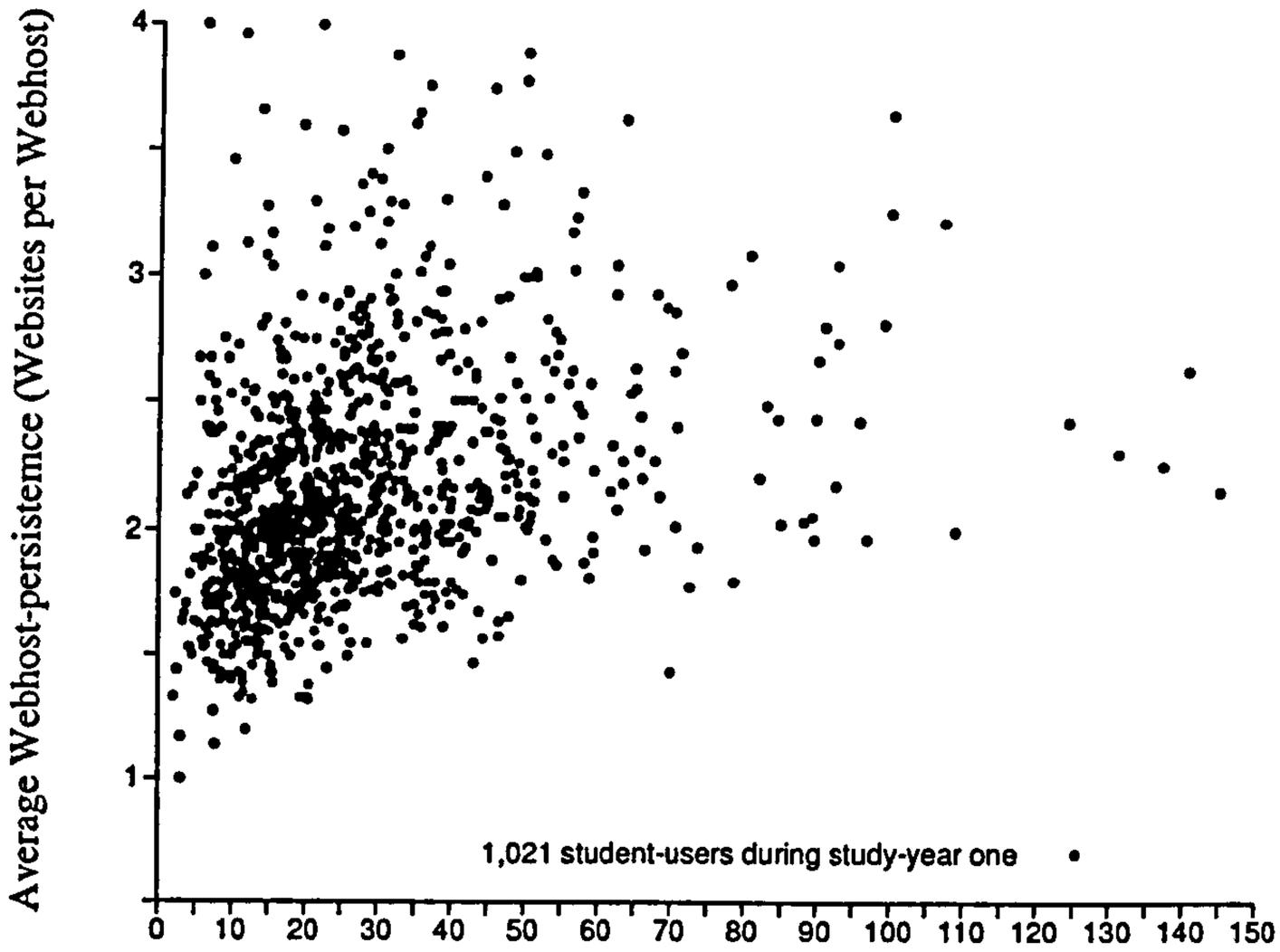

Figure C.6: Scattergram of 1,021 student-user's average session click rate and average Webhost-persistence during study-year one (range illustrated up to 150 clicks per session and four Websites per Webhost)

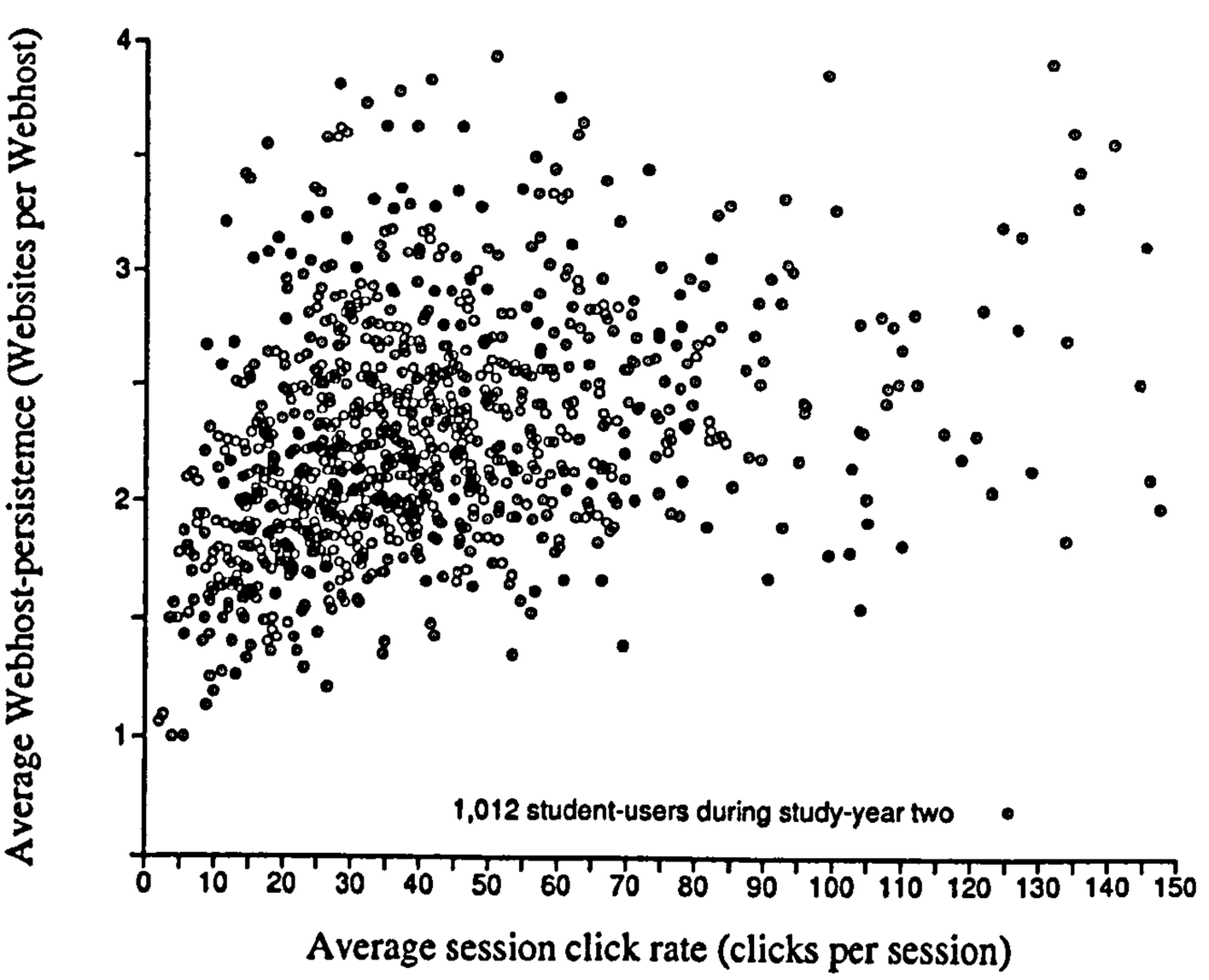

Figure C.7: Scattergram of 1,012 student-user's average session click rate and average Webhost-persistence during study-year two (range illustrated up to 150 clicks per session and four Websites per Webhost)

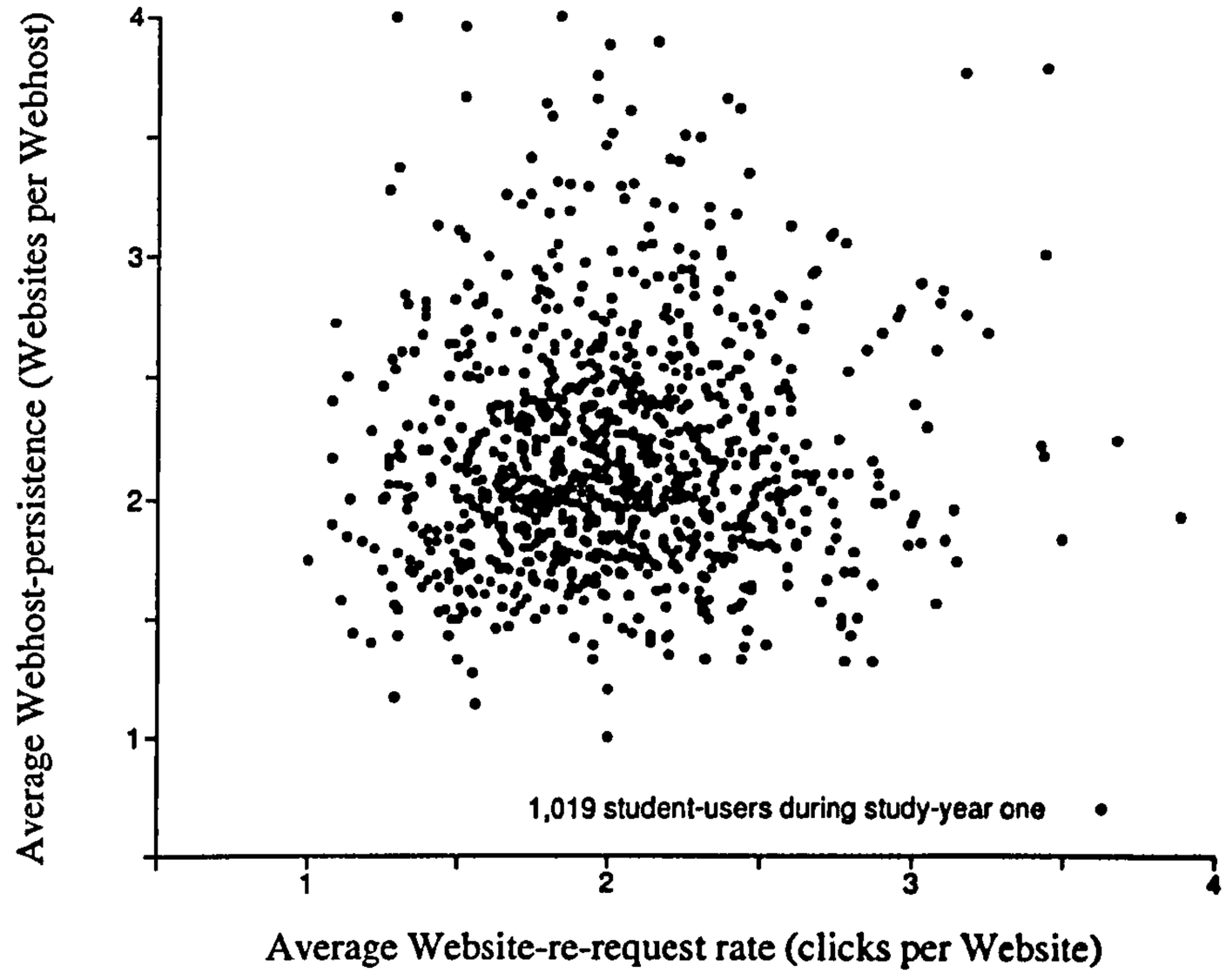

Figure C.8: Scattergram of 1,019 student-user's average Website-re-request rate and average Webhost-persistence during study-year one (range illustrated up to four clicks per Website and four Websites per Webhost

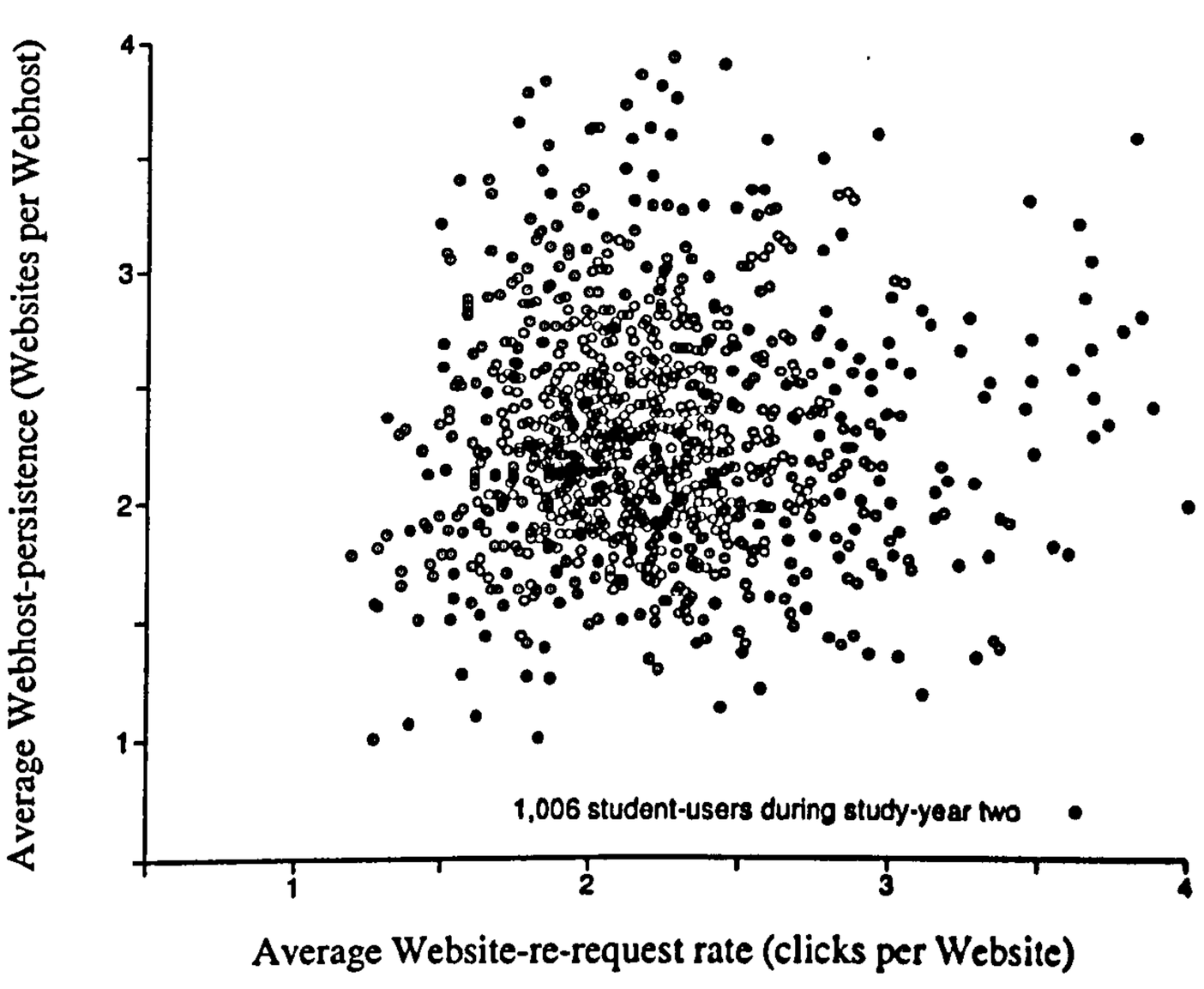

Figure C.9: Scattergram of 1,006 student-user's average Website-re-request rate and average Webhost-persistence during study-year two (range illustrated up to four clicks per Website and four Websites per Webhost

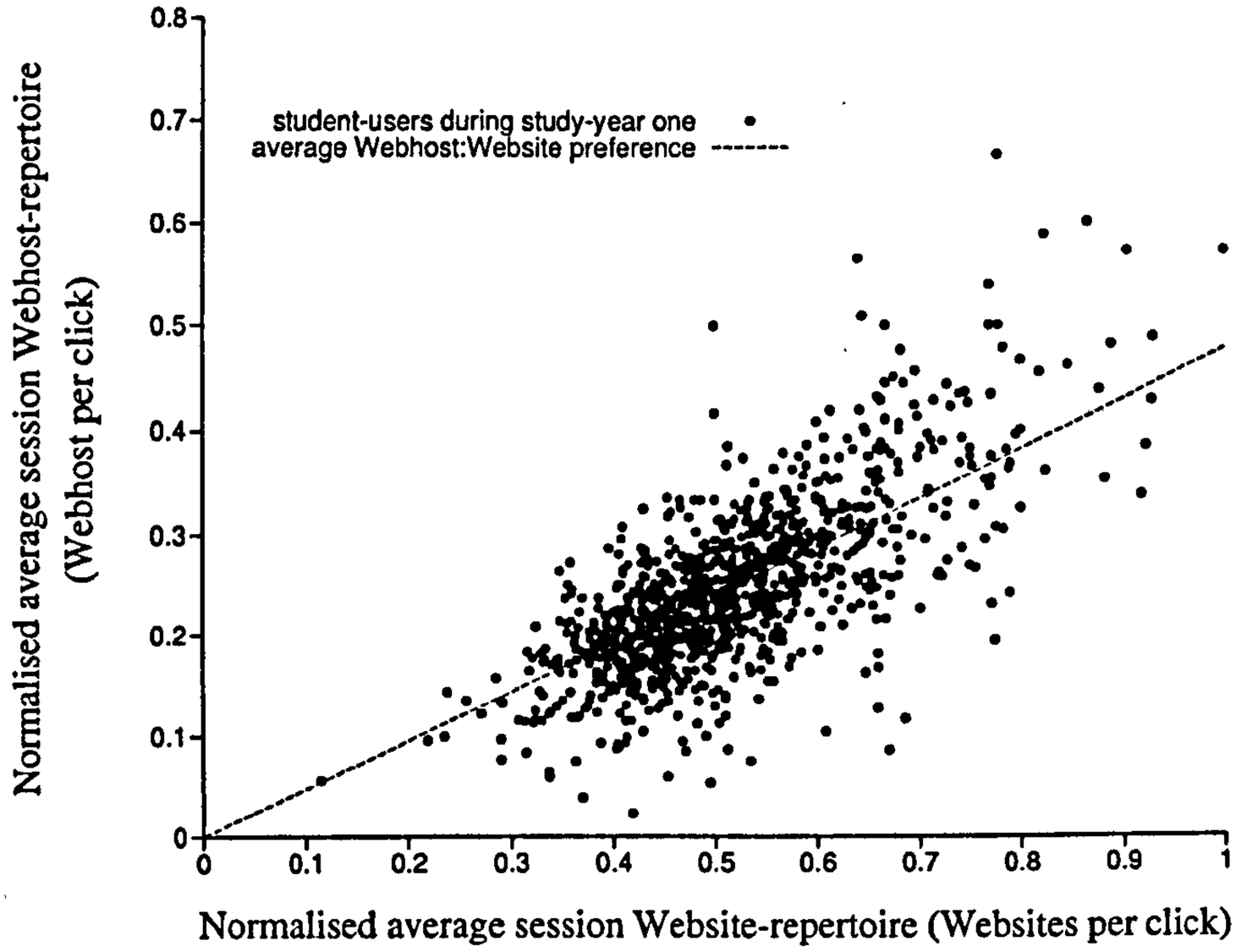

Figure C.10: Scattergram of student-user's click normalised average session Website-repertoire and average session Webhost-repertoires during study-year one

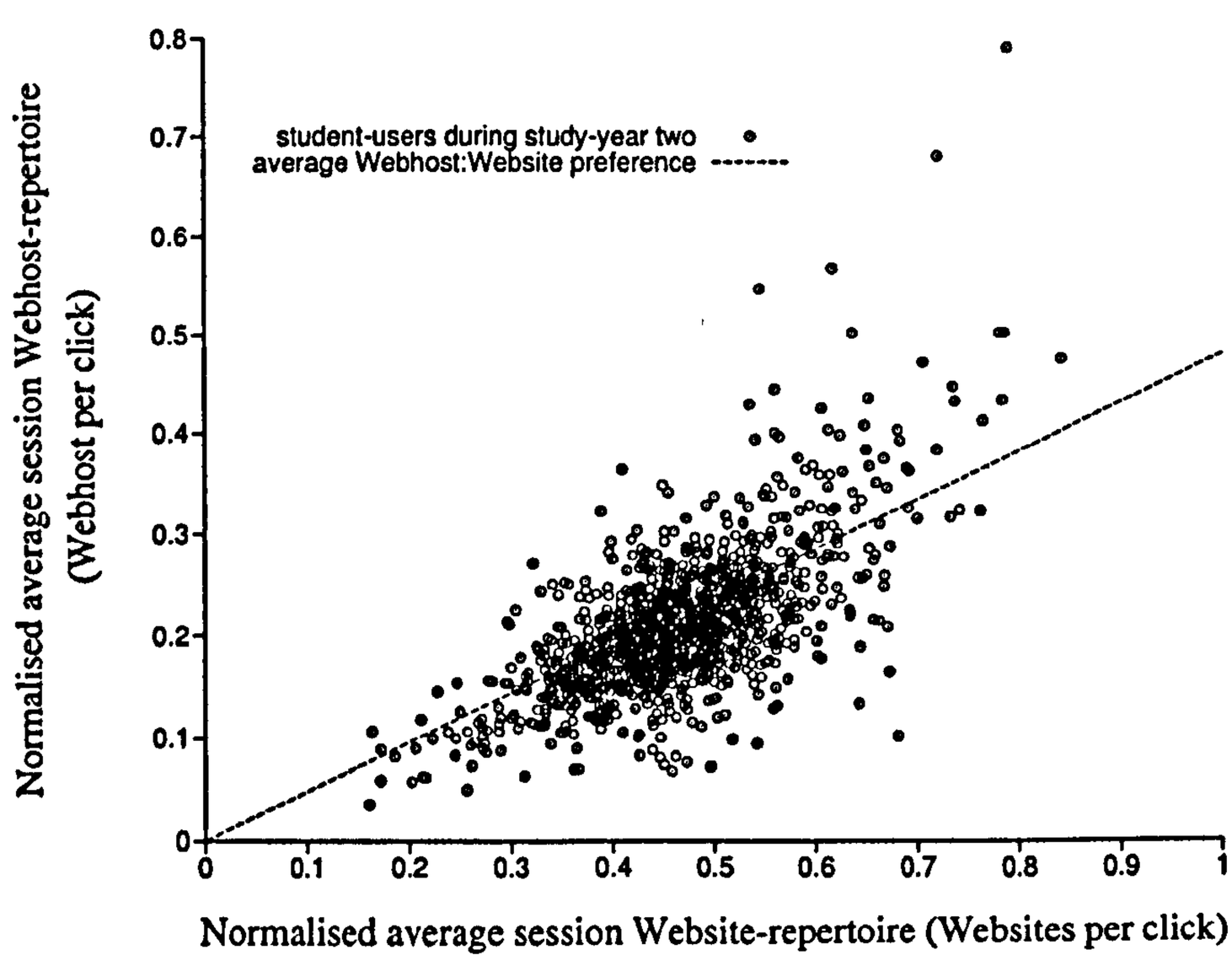

Figure C.11: Scattergram of student-user's click normalised average session Website-repertoire and average session Webhost-repertoires during study-year two

Average session-conformance and session-conformance range

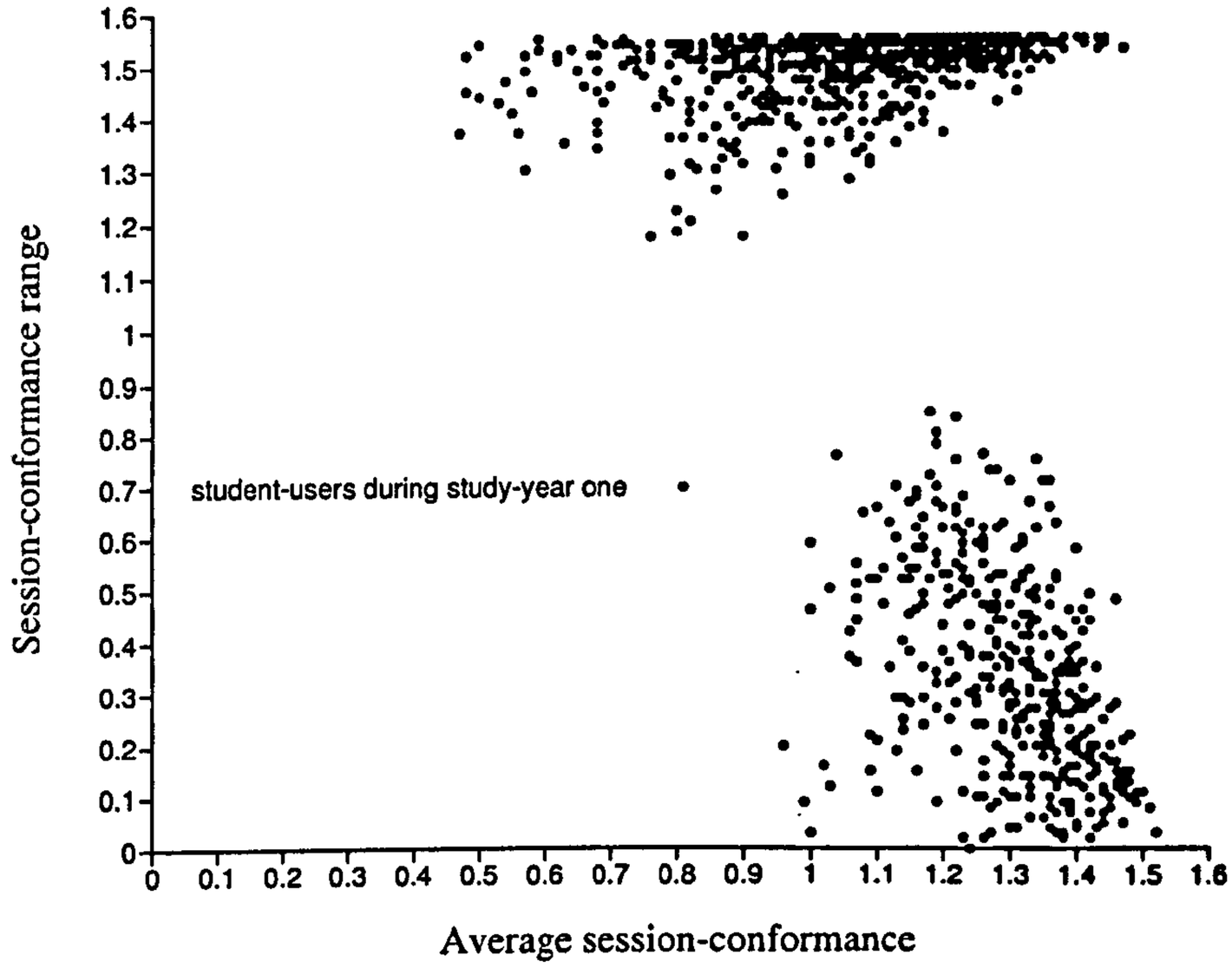

Figure C.12: Scattergram of student-user's average session-conformance and session-conformance range during study-year one

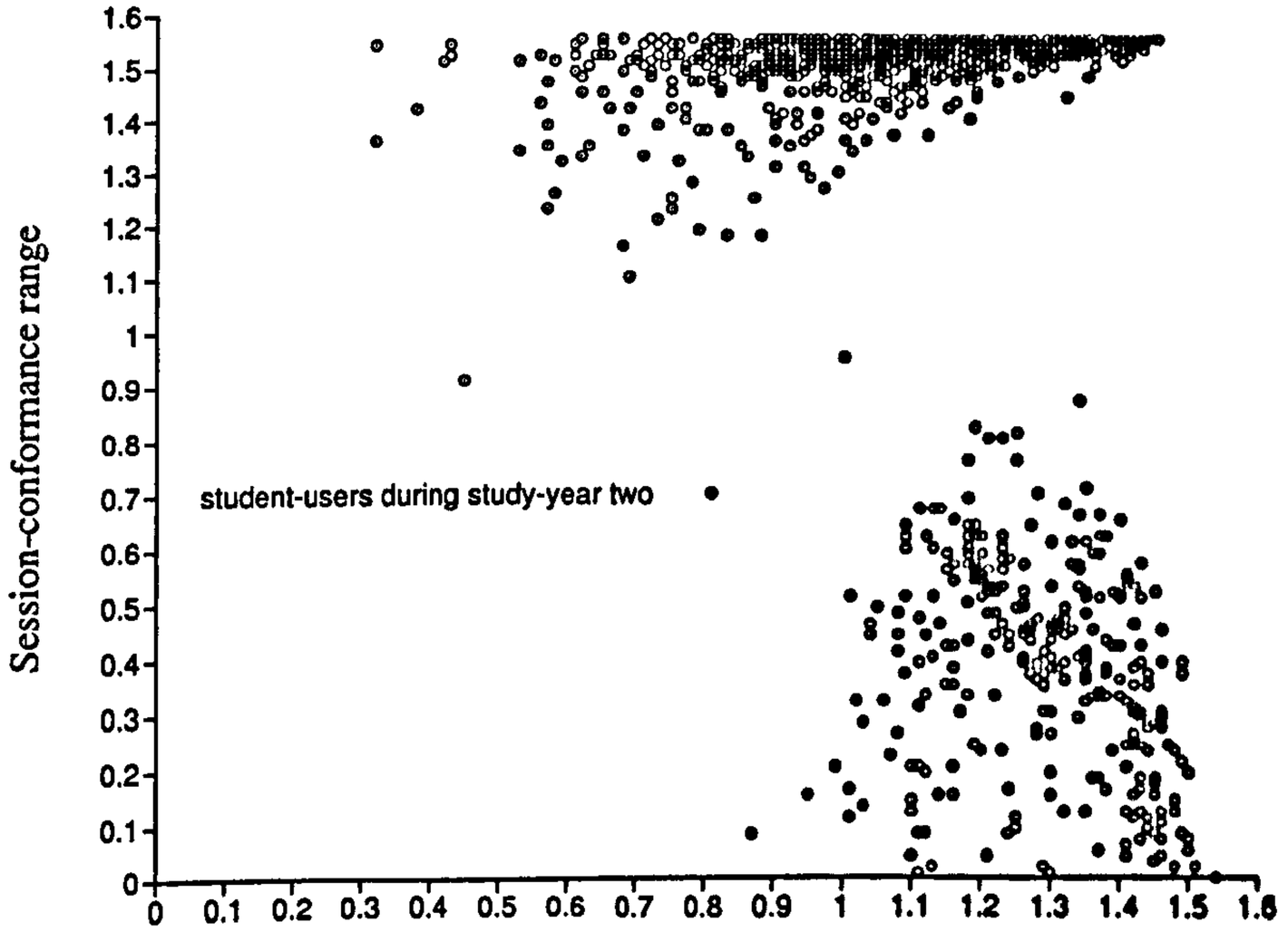

Figure C.13: Scattergram of student-user's average session-conformance and session-conformance range during study-year two

Similarities and differences between attribute group membership

User-attribute	User-attribute			
	smaller	larger	conformant	eclectic
study-year one				
men	309	231	198	342
women	405	105	197	313
smaller			352	362
larger			43	293
study-year two				
men	308	232	151	389
women	361	149	143	367
smaller			253	416
larger			41	340

Table C.3: Cross-tabulation of student-user frequencies by user-attributute

Similarities and differences between study-years one and two

Student-users	User-characterization						
	session-by-session					session-to-session	
	average session click rate (clks/session)	average query-click proportion (qry-clks/clk)	average Website-re-request rate (clks/Website)	average Webhost-persistence (Wsites/Whost)	average session-conformance	Website-trajectory slope (Wsites/clk)	session-conformance range
gender							
540 men	32.4(1.42)	0.25(0.005)	2.02(0.02)	2.25(0.03)	1.16(0.008)	0.39(0.005)	1.08(0.02)
510 women	25.3(0.86)	0.26(0.005)	2.03(0.02)	2.29(0.05)	1.14(0.010)	0.42(0.006)	1.04(0.03)
session-rate							
714 smaller	25.9(0.70)	0.26(0.004)	2.00(0.02)	2.23(0.03)	1.16(0.007)	0.43(0.005)	0.90(0.02)
336 larger	35.5(2.15)	0.24(0.005)	2.08(0.03)	2.35(0.06)	1.12(0.011)	0.35(0.005)	1.39(0.02)
conformance							
395 conformant	29.4(1.14)	0.28(0.006)	2.05(0.02)	2.24(0.06)	1.30(0.006)	0.42(0.008)	0.34(0.01)
655 eclectic	28.7(1.18)	0.24(0.004)	2.01(0.02)	2.29(0.03)	1.06(0.007)	0.40(0.005)	1.49(0.003)

Table C.4: Cross-tabulation of mean user-characterization metric by user-attribute partition and user-characterization during study-year one

Student-users	User-characterization						
	session-by-session					session-to-session	
	average session click rate (clks/session)	average query-click proportion (qry-clks/clk)	average Website-re-request rate (clks/Website)	average Webhost-persistence (Wsites/Whost)	average session-conformance	Website-trajectory slope (Wsites/clk)	session-conformance range
gender							
540 men	50.0(1.89)	0.31(0.005)	2.21(0.02)	2.37(0.03)	1.11(0.009)	0.38(0.004)	1.19(0.02)
510 women	37.9(1.02)	0.34(0.006)	2.30(0.03)	2.35(0.03)	1.11(0.010)	0.37(0.005)	1.19(0.02)
session-rate							
669 smaller	41.8(1.38)	0.34(0.005)	2.28(0.02)	2.32(0.03)	1.10(0.008)	0.39(0.004)	1.06(0.02)
381 larger	48.2(1.83)	0.31(0.005)	2.20(0.02)	2.43(0.03)	1.12(0.011)	0.35(0.004)	1.41(0.02)
conformance							
294 conformant	48.4(2.16)	0.34(0.008)	2.34(0.04)	2.31(0.05)	1.28(0.008)	0.37(0.007)	0.38(0.01)
756 eclectic	42.5(1.28)	0.32(0.004)	2.21(0.02)	2.38(0.02)	1.04(0.007)	0.38(0.004)	1.50(0.002)

Table C.5: Cross-tabulation of mean user-characterization metric by user-attribute partition and user-characterization during study-year two

Website popularity

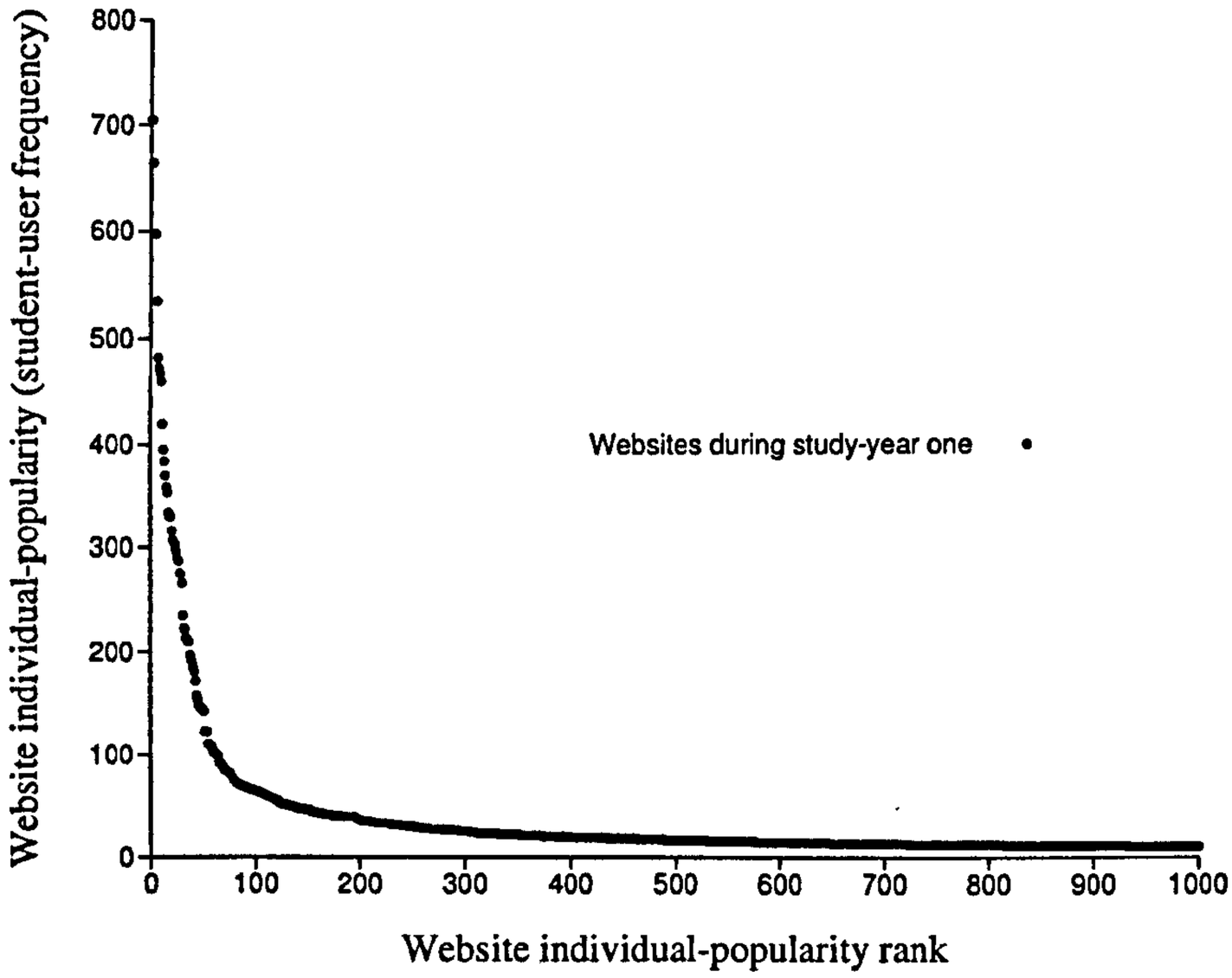

Figure C.14: Zipf distribution of Website individual-popularity during study-year one

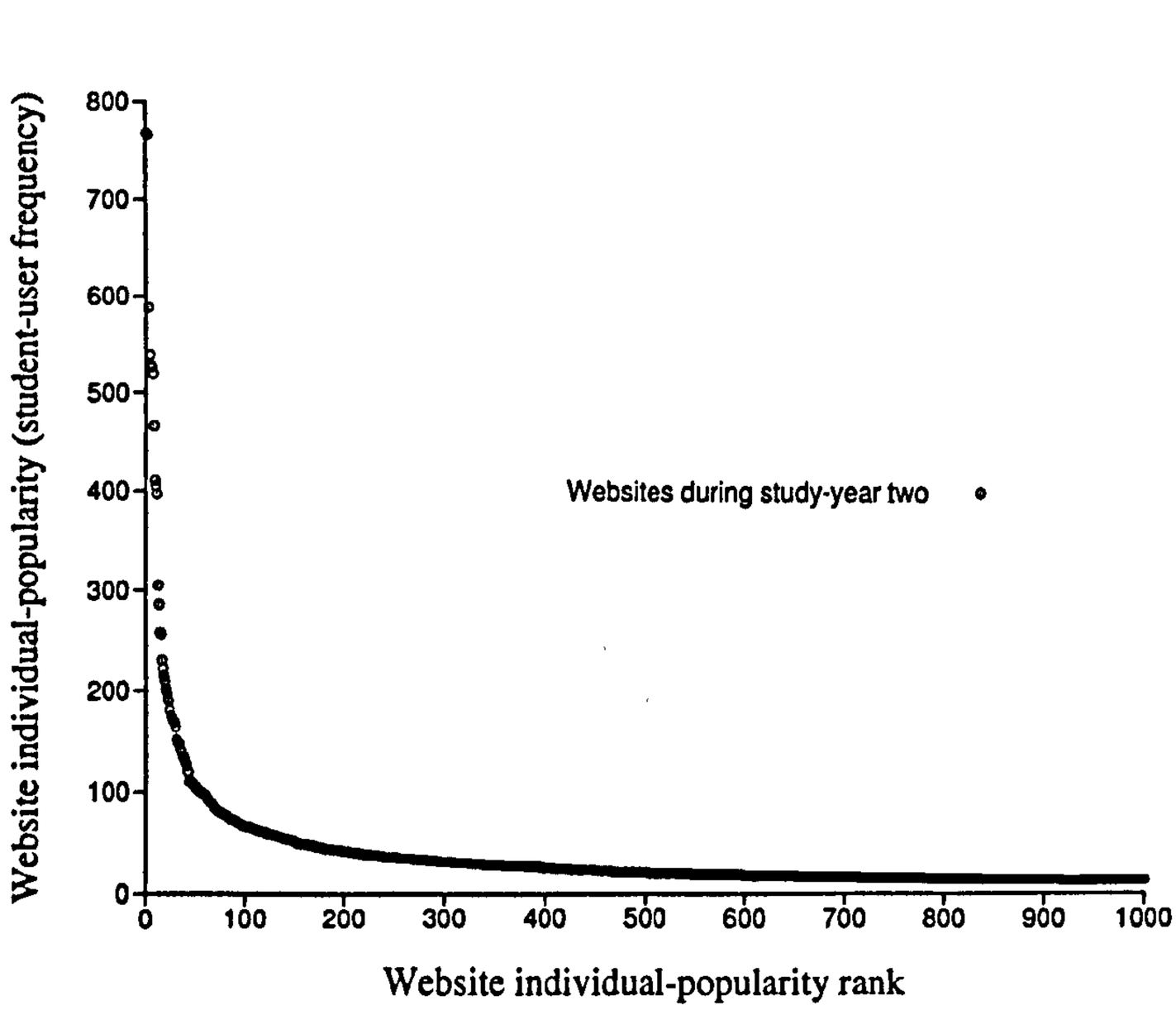

Figure C.15: Zipf distribution of Website individual-popularity during study-year two

Rank	Proportion of student-users	Website (conditioned url-string)
1	67.1%	<www.yahoo.akadns.net/>
2	63.2%	<www-uk.netscape.com/intl_search/uk/escapes/internet_search.html>
3=	57.0%	<search.snv.yahoo.com/bin/search>
3=	57.0%	<www-uk.netscape.com/escapes/search/choose.html>
5	56.9%	<www-uk.netscape.com/home/internet-search.html>
6	51.0%	<homerc.europe.yahoo.com/>
7	45.9%	<www-uk.netscape.com/uk/escapes/search/ntsrchrnd-*.html>
8	45.0%	<www-uk.netscape.com/bookmark/4_5/tsearch.html>
9	44.8%	<search.yahoo.co.uk/search/ukie>
10	43.9%	<www-uk.netscape.com/escapes/internet_search.html>
11	40.0%	<www-uk.netscape.com/uk/escapes/internet_search.html>
12	37.6%	<google.yahoo.com/bin/query>
13	36.6%	<nodsearch.netscape.com/search.gw>
14	35.2%	<search.snv.yahoo.com/search>
15	34.2%	<altavista.com/cgi-bin/query>
16	33.7%	<www-uk.netscape.com/>
17	31.8%	<www.geocities.com/ad_container/pop.html>
18=	31.4%	<www.infoseek.com/Titles>
18=	31.4%	<uk.excite.co.uk/search.gw>
20	30.1%	<google.yahoo.com/bin/query_uk>

Table C.6: Relative-individual-popularity of 'top-twenty' Websites during study-year one

Rank	Proportion of student-users	Website (conditioned url-string)
1	73.1%	<utility3-search.europe.yahoo.com/search/ukie>
2	73.0%	<home.europe.yahoo.com/>
3	56.0%	<google.yahoo.akadns.net/bin/query_uk>
4	51.3%	<homerc.europe.yahoo.com/>
5	50.4%	<altavista.com/cgi-bin/query>
6	50.1%	<www.altavista.magallanes.net/cgi-bin/query>
7	49.4%	<altavista.com/cgi-bin/query>
8	44.4%	<www.yahoo.akadns.net/>
9	39.1%	<www.infoseek.com/Home>
10	38.6%	<www.infoseek.com/Titles>
11	37.8%	<search.snv.yahoo.com/bin/search>
12	29.0%	<search.yahoo.co.uk/search/ukie>
13	27.2%	<google.yahoo.com/bin/query>
14	24.6%	<cgi.netscape.com/cgi-bin/plug-in_finder.cgi>
15	24.4%	<search.snv.yahoo.com/search>
16	21.9%	<netscape.google.com/netscape>
17	21.1%	<www.savvysearch.com/>
18	20.5%	<members.tripod.com/adm/popup/roadmap.shtml>
19	20.0%	<www.altavista.magallanes.net/av/eng/help.htm>
20	19.2%	<search.netscape.com/cgi-bin/search>

Table C.7: Relative-individual-popularity of 'top-twenty' Websites during study-year two

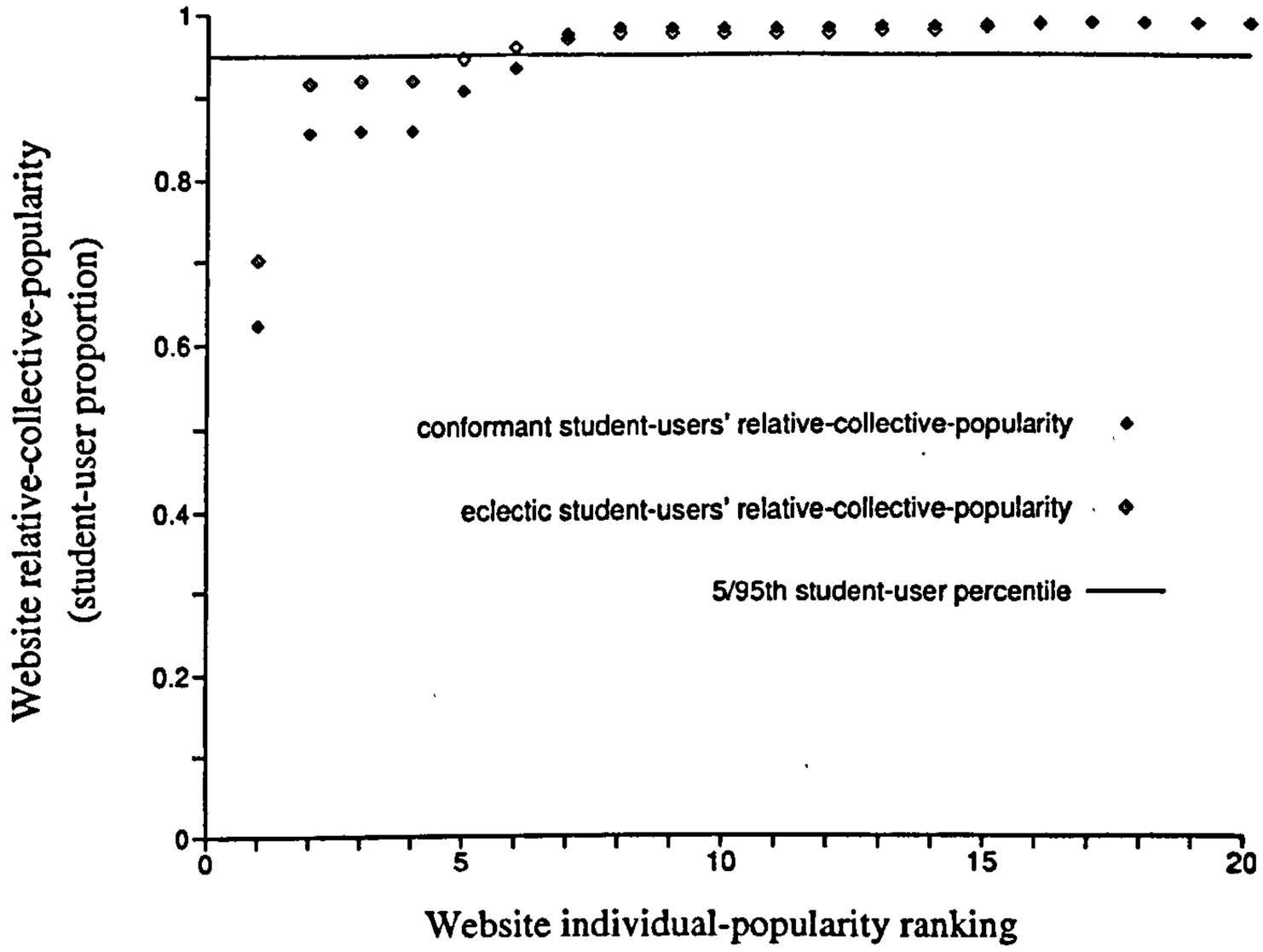

Figure C.16: Ranked distributions of Website relative-collective-popularity by Website individual-popularity for conformant and eclectic student-users during study-year one

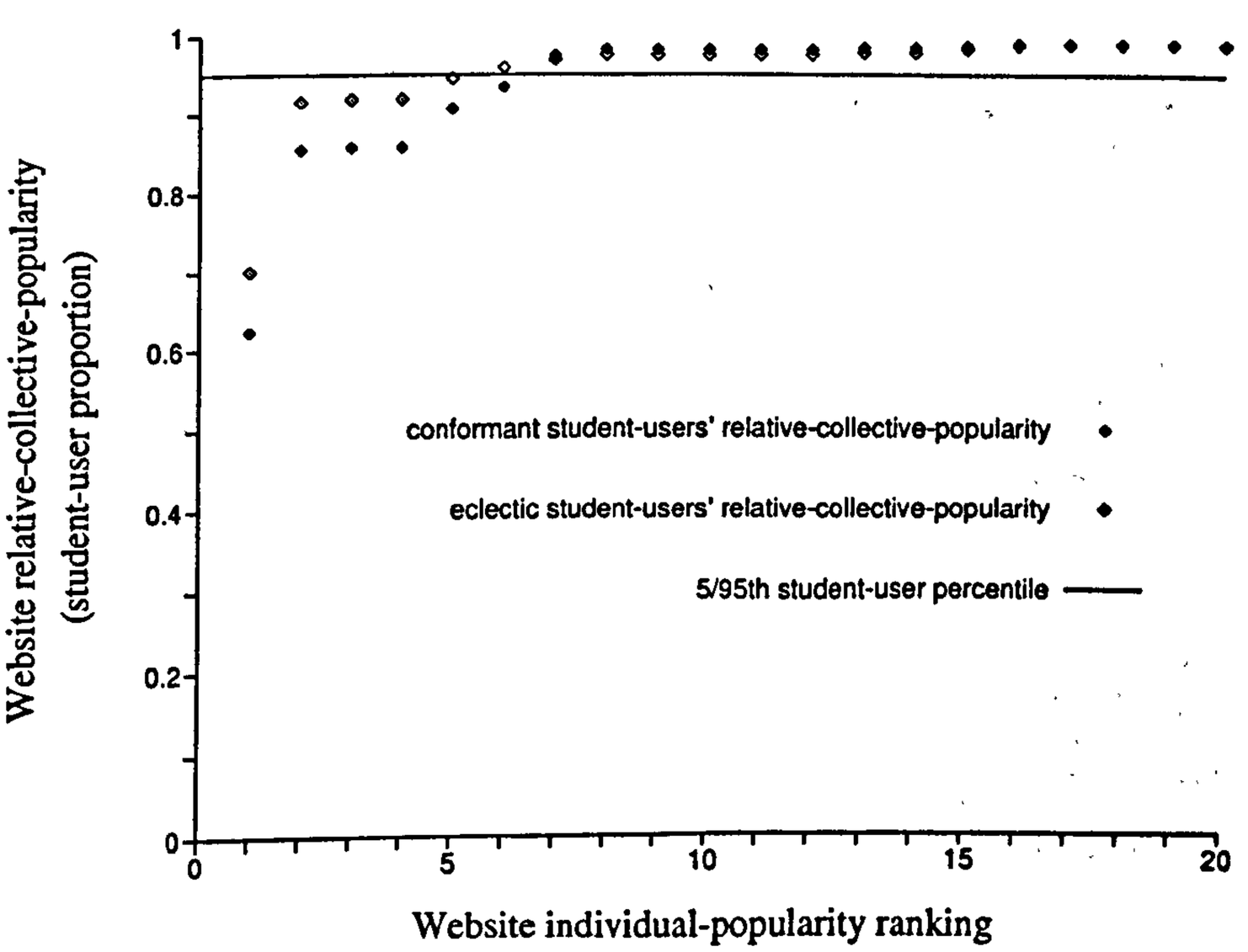

Figure C.17: Ranked distributions of Website relative-collective-popularity by Website individual-popularity for conformant and eclectic student-users during study-year two

Appendix D

How do student-users use Web information location services?

Web information location service usage

Tables D.1 and D.2 report the session frequencies which are discussed in Chapter five.

Web search-query analyses

Analyses of the AlataVista-Excite sample of 'search-engine' search-queries are discussed in Chapter five. Table D.3 reports the student-user frequencies not included in the sample while Tables D.4 and D.5 report search-user frequencies.

Figures D.1 and D.2 illustrate scattergrams which relate the number of search-terms in a search-query to the number of search-queries in a search-session.

Tables D.6 to D.8 report search-user frequencies according to the user's average search-query and average search-term counts. Tables D.9 to D.11 report frequencies of search-queries while Tables D.12 and D.13 report search-user frequencies.

Web information location service usage

User-attribute	Sessions	
	search-sessions	not search-sessions
gender		
men	7,140 sessions	6,763 sessions
women	3,343 sessions	4,120 sessions
session-rate		
smaller	3,548 sessions	3,130 sessions
larger	6,935 sessions	7,753 sessions
conformance		
conformant	2,541 sessions	1,697 sessions
eclectic	7,942 sessions	9,186 sessions
Study-year two	10,483 sessions	10,883 sessions

Table D.1: Cross-tabulation of session frequency by user-attribute partition and 'searching' during study-year one

User-attribute	Sessions	
	search-sessions	not search-sessions
gender		
men	7,858 sessions	7,365 sessions
women	4,453 sessions	5,516 sessions
session-rate		
smaller	3,931 sessions	3,422 sessions
larger	8,380 sessions	9,459 sessions
conformance		
conformant	2,288 sessions	1,875 sessions
eclectic	10,023 sessions	11,006 sessions
Study-year two	12,311 sessions	12,881 sessions

Table D.2: Cross-tabulation of session frequency by user-attribute partition and 'searching' during study-year two

The AltaVista-Excite sample

	Gender	
	men	women
not in the AltaVista-Excite sample all student-users	55 student-users 540 student-users	69 student-users 510 student-users

	Session-rate	
	smaller	larger
Study-year one not in the AltaVista-Excite sample all student-users	105 student-users 714 student-users	19 student-users 336 student-users
Study-year two not in the AltaVista-Excite sample all student-users	109 student-users 669 student-users	15 student-users 381 student-users

	Conformance	
	eclectic	conformant
Study-year one not in the AltaVista-Excite sample all student-users	61 student-users 655 student-users	63 student-users 395 student-users
Study-year two not in the AltaVista-Excite sample all student-users	78 student-users 756 student-users	46 student-users 294 student-users

Table D.3: Cross-tabulation of AltaVista-Excite sample complement student-user frequencies by user-attribute partition

Search-user's search-queries

User-attribute	Average search-query count search-queries per search-session	
	< 3	⩾ 3
gender		
men	279 search-users	135 search-users
women	241 search-users	101 search-users
session-rate		
smaller	330 search-users	136 search-users
larger	190 search-users	100 search-users
conformance		
conformant	189 search-users	77 search-users
eclectic	331 search-users	159 search-users

Table D.4: Cross-tabulation of AltaVista-Excite search-user frequencies by user-attribute partition and average search-query count during study-year one

User-attribute	Average search-query count search-queries per search-session	
	< 3	⩾ 3
gender		
men	245 search-users	162 student-users
women	225 search-users	122 student-users
session-rate		
smaller	261 search-users	162 student-users
larger	209 search-users	122 student-users
conformance		
conformant	106 search-users	63 search-users
eclectic	364 search-users	221 search-users

Table D.5: Cross-tabulation of AltaVista-Excite search-user frequencies by user-attribute partition and average search-query count during study-year two

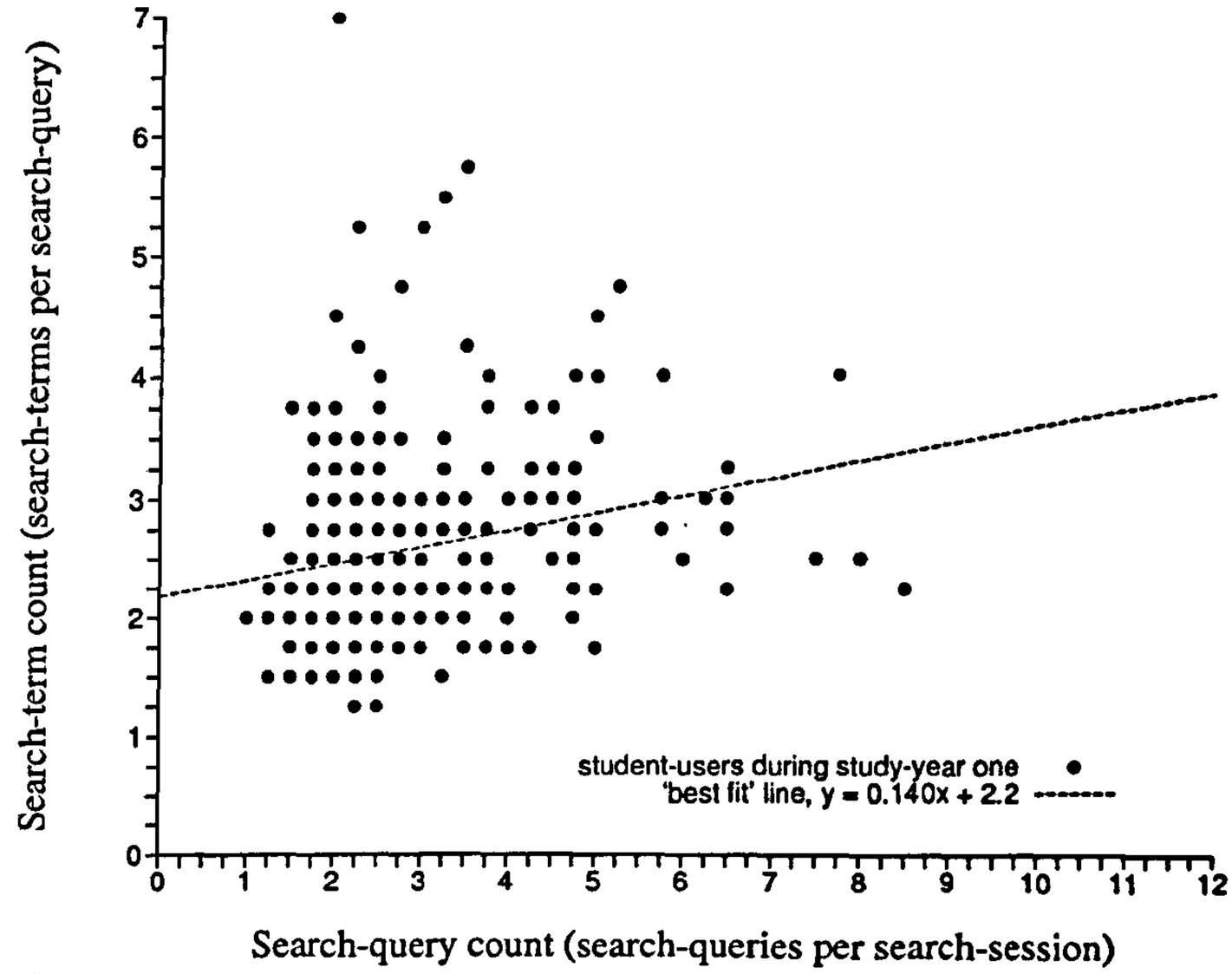

Figure D.1: Scattergram of 258 search-user's average search-query count and search-term count during study-year one

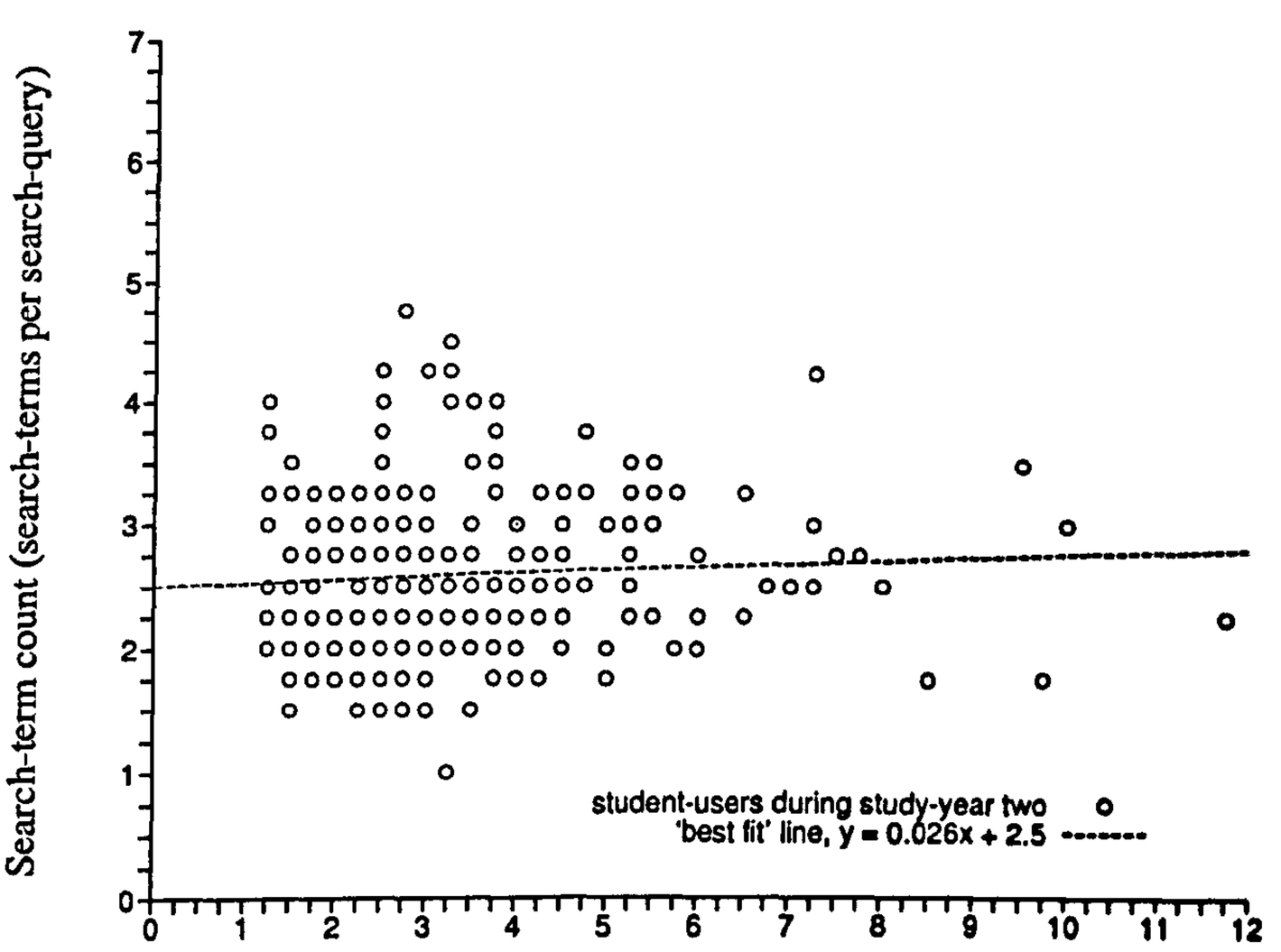

Figure D.2: Scattergram of 292 search-user's average search-query count and search-term count during study-year two

Average search-query count search-queries per search-session	**Average search-term count** search-terms per search-query	
	< 3	$\geqslant 3$
study-year one < 3 $\geqslant 3$	418 search-users 152 search-users	102 search-users 84 search-users
study-year two < 3 $\geqslant 3$	345 search-users 206 search-users	125 search-users 78 search-users

Table D.6: Cross-tabulation of AltaVista-Excite search-user frequency by average search-query count and average search-term count

User-attribute	Average search-term count search-terms per search-query	
	< 3	⩾ 3
gender		
men	323 search-users	91 search-users
women	247 search-users	95 search-users
session-rate		
smaller	347 search-users	119 search-users
larger	223 search-users	67 search-users
conformance		
conformant	198 search-users	68 search-users
eclectic	372 search-users	118 search-users

Table D.7: Cross-tabulation of AltaVista-Excite search-user frequency by user-attribute partition and average search-term count during study-year one

User-attribute	Average search-term count search-terms per search-query	
	< 3	⩾ 3
gender		
men	308 search-users	99 search-users
women	243 search-users	104 search-users
session-rate		
smaller	315 search-users	108 search-users
larger	236 search-users	95 search-users
conformance		
conformant	123 search-users	46 search-users
eclectic	428 search-users	157 search-users

Table D.8: Cross-tabulation of AltaVista-Excite search-user frequency by user-attribute partition and average search-term count during study-year two

AltaVista and Excite search-queries and search-terms

Sessions	Search-queries	
	singleton	non-singleton
study-year one smaller larger	1,057 search-queries 2,101 search-queries	2,429 search-queries 7,613 search-queries
study-year two smaller larger	908 search-queries 2,456 search-queries	2,419 search-queries 9,328 search-queries

Table D.9: Cross-tabulation of search-query frequency by search-session size and search-query count

User-attribute	Search-queries	
	singleton	non-singleton
gender		
men	2,246 search-queries	7,003 search-queries
women	912 search-queries	3,039 search-queries
session-rate		
smaller	857 search-queries	2,444 search-queries
larger	2,301 search-queries	7,598 search-queries
conformance		
conformant	619 search-queries	1,781 search-queries
eclectic	2,539 search-queries	8,261 search-queries

Table D.10: Cross-tabulation of search-query frequency by user-attribute partition and search-query count during study-year one

User-attribute	Search-queries	
	singleton	non-singleton
gender		
men	2,265 search-queries	7,977 search-queries
women	1,099 search-queries	3,770 search-queries
session-rate		
smaller	1,112 search-queries	3,450 search-queries
larger	2,252 search-queries	8,297 search-queries
conformance		
conformant	529 search-queries	1,850 search-queries
eclectic	2,835 search-queries	9,897 search-queries

Table D.11: Cross-tabulation of search-query frequency by user-attribute partition and search-query count during study-year two

User-attribute	Average search-term count	
	singleton	> one search-term per search-query
gender		
men	28 search-users	386 search-users
women	36 search-users	306 search-users
session-rate		
smaller	54 search-users	412 search-users
larger	10 search-users	280 search-users
conformance		
conformant	31 search-users	235 search-users
eclectic	33 search-users	457 search-users

Table D.12: Cross-tabulation of AltaVista-Excite search-user frequency by user-attribute partition and average search-term count during study-year one

User-attribute	Average search-term count	
	singleton	> one search-term per search-query
gender		
men	32 search-users	375 search-users
women	18 search-users	329 search-users
session-rate		
smaller	34 search-users	389 search-users
larger	16 search-users	315 search-users
conformance		
conformant	13 search-users	156 search-users
eclectic	37 search-users	548 search-users

Table D.13: Cross-tabulation of AltaVista-Excite search-user frequency by user-attribute partition and average search-term count during study-year two

Appendix E

How do novices seek Web information?

How do student-users change their Web information seeking activity?

Tables E.1 to E.8 illustrate conditional analyses of the average session click rate and average session-conformance user characterizations all of which appear to show a novice-effect. Tables E.9 to E.24 illustrate conditional analyses of the average query-click proportion, average Website-re-request rate, average Webhost-persistence and Website-trajectory slope user characterizations none of which show a novice-effect. The analyses include by-cohort, by-gender and by-joint-session-rate. Conditional-analysis is discussed in Chapter three and the novice-effect is discussed in Chapter six.

How do search-users change their use of Web information location services?

Tables E.25 to E.32 illustrate conditional analyses of the average search-query proportion and search-session proportion user characterizations of how search-users use Web information location services. These conditional analyses are in respect of the 1,002 search-users during either study-year one or two. The 1997/1998 cohort, men/women gender and smaller/larger joint-session-rate partitions are 392/610, 529/473 and 640/362 student-users respectively.

Tables E.33 to E.40 illustrate conditional analyses of the average search-query count and average search-term count in respect of the 584 search-users in the AltaVista-Excite sample during each of study-years one and two. The 1997/1998 cohort, men/women gender and smaller/larger joint-session-rate partitions are 220/364, 336/248 and 306/278 student-users respectively.

average session click rate

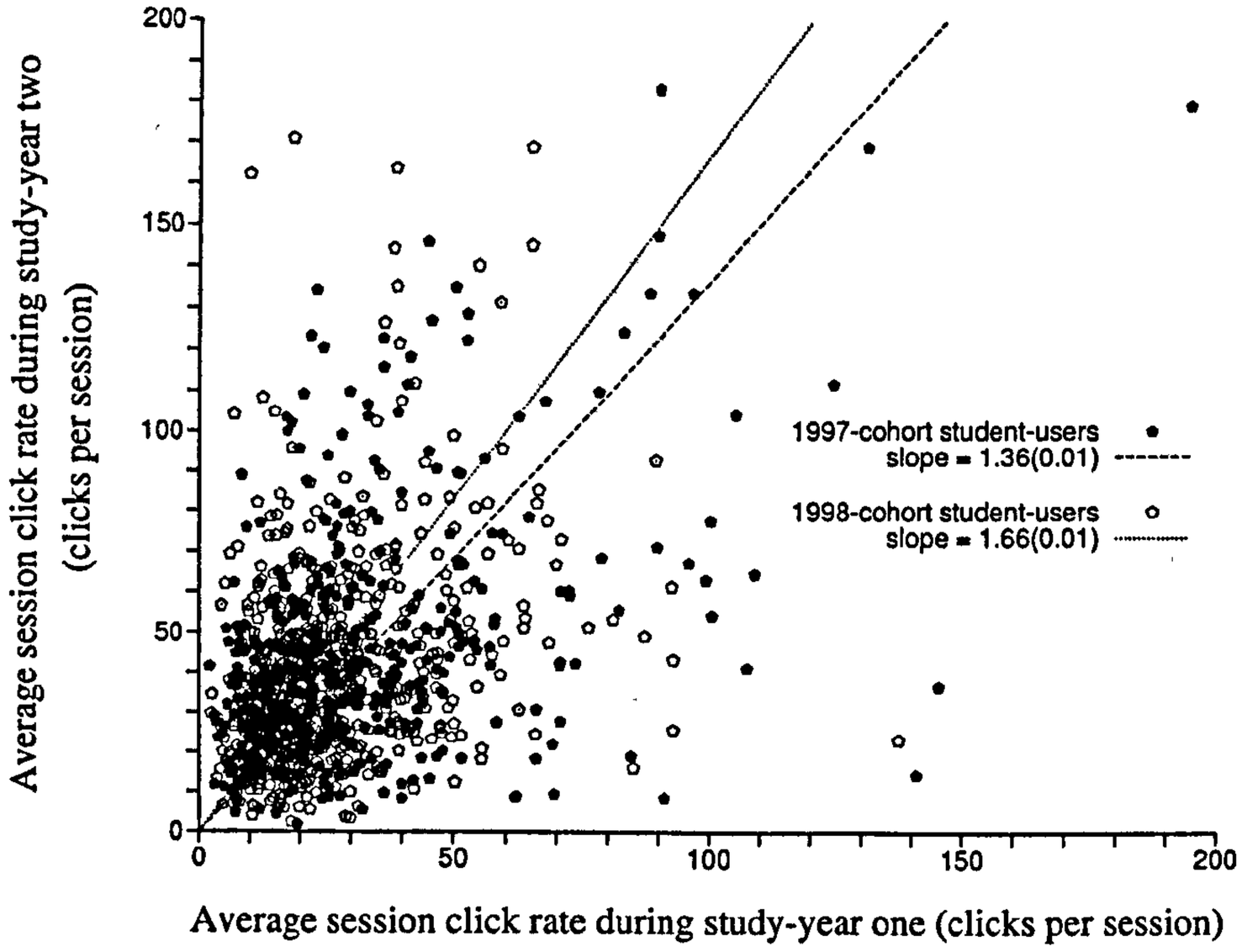

Figure E.1: Conditional distributions of student-user's average session click rate by-cohort (range illustrated up to 200 clicks per session)

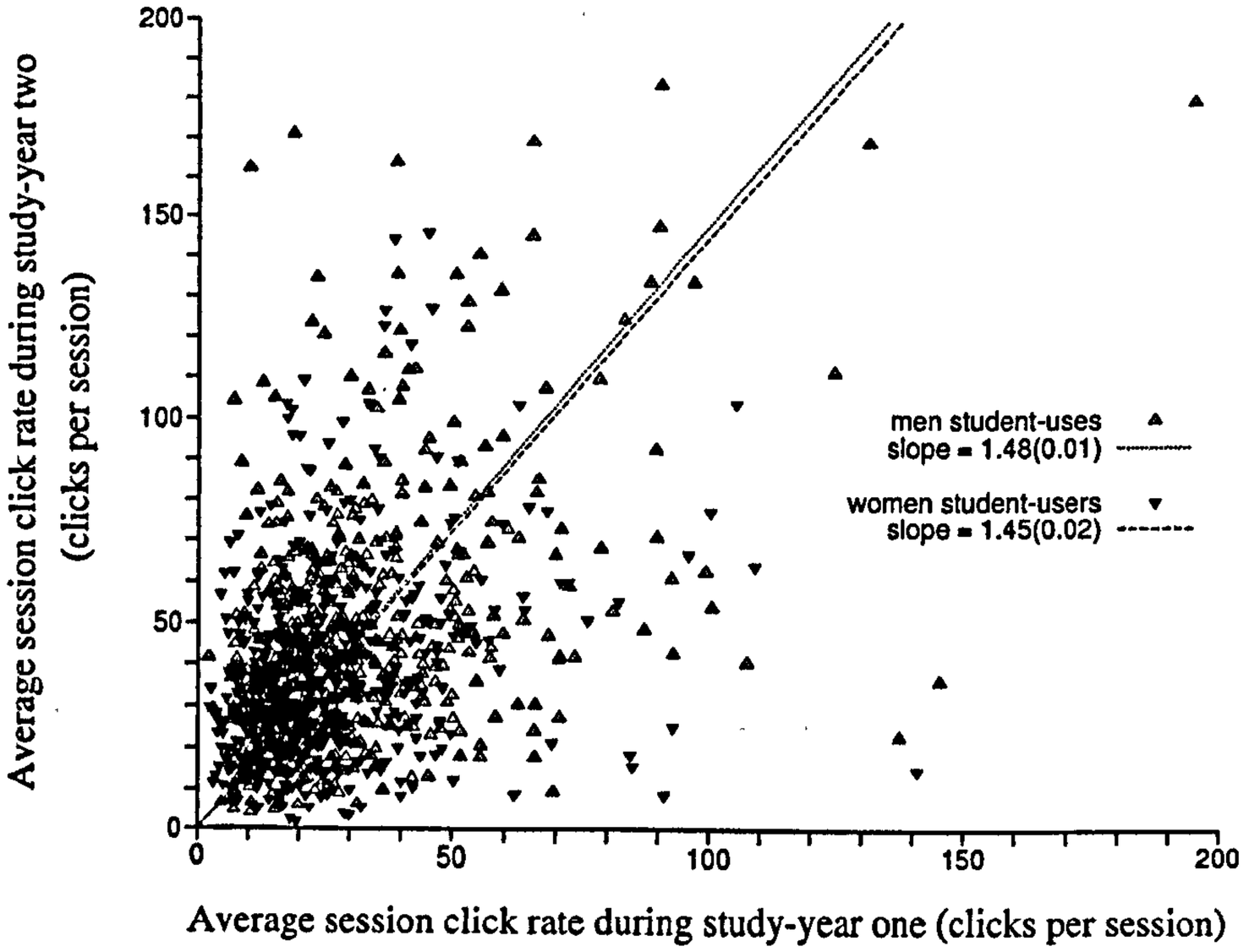

Figure E.2: Conditional distributions of student-user's average session click rate by-gender (range illustrated up to 200 clicks per session)

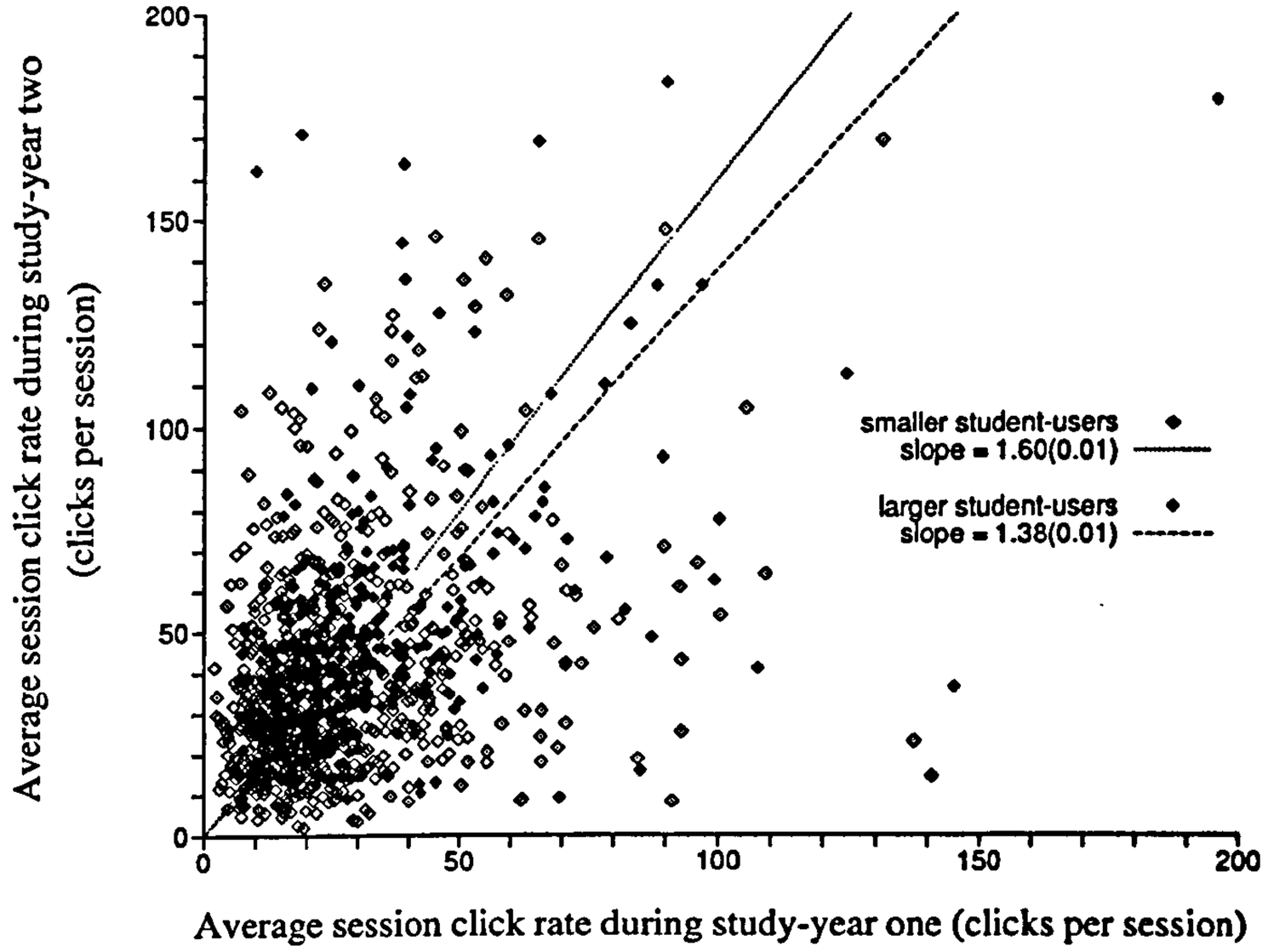

Figure E.3: Conditional distributions of student-user's average session click rate by-joint-session-rate (range illustrated up to 200 clicks per session)

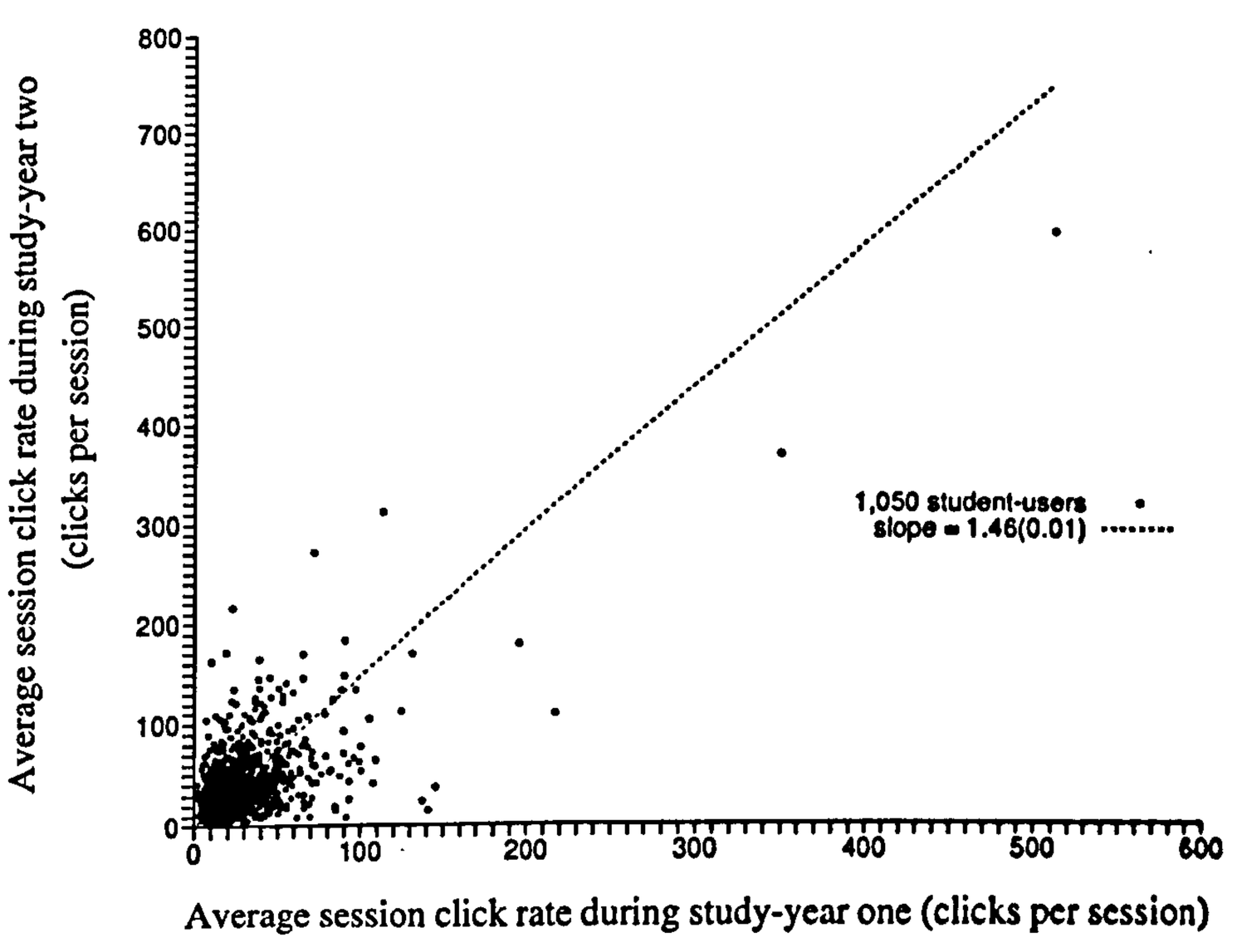

Figure E.4: Conditional distribution of student-user's average session click rate

average session-conformance

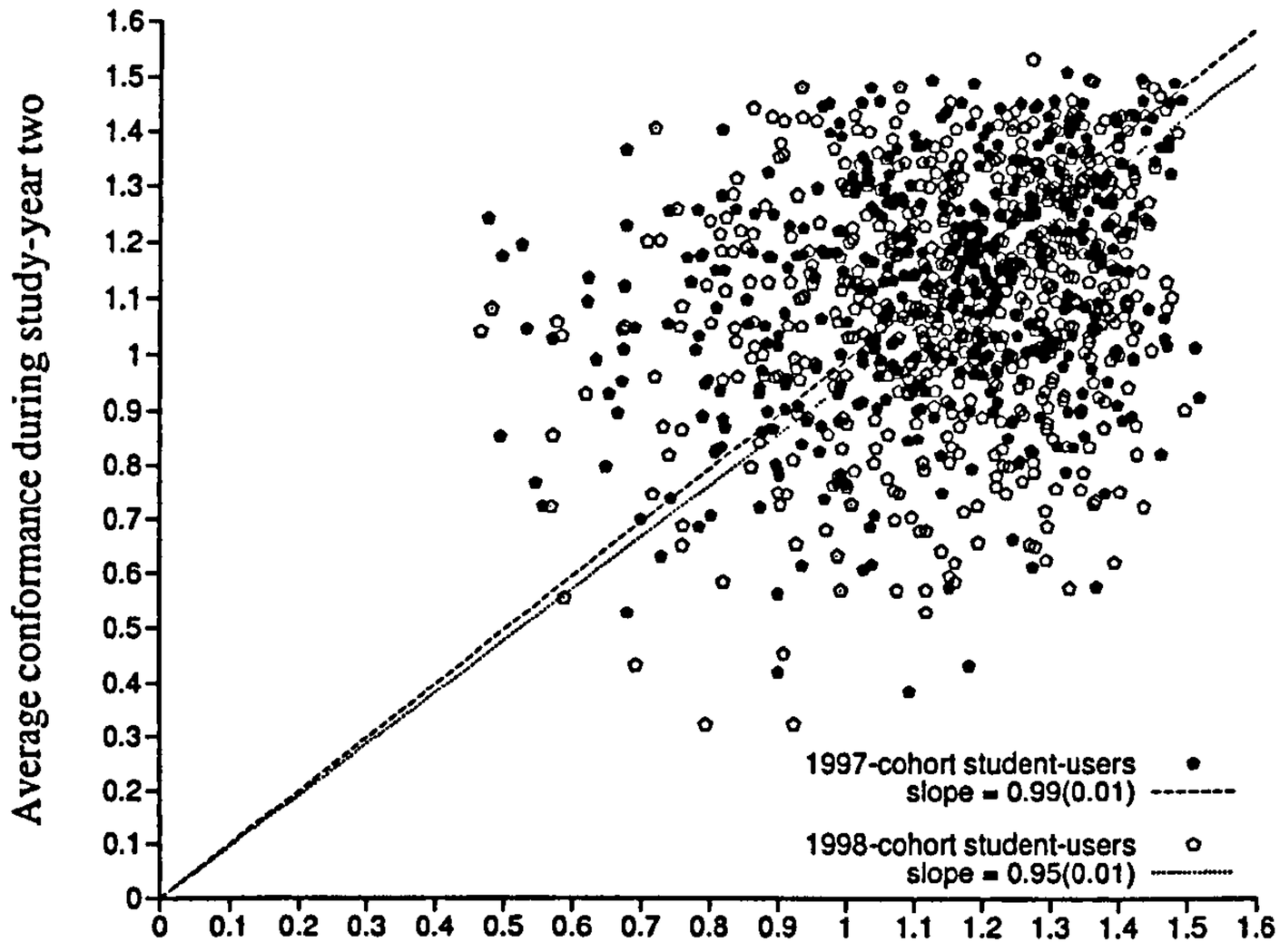

Figure E.5: Conditional distributions of student-user's average session-conformance by-cohort

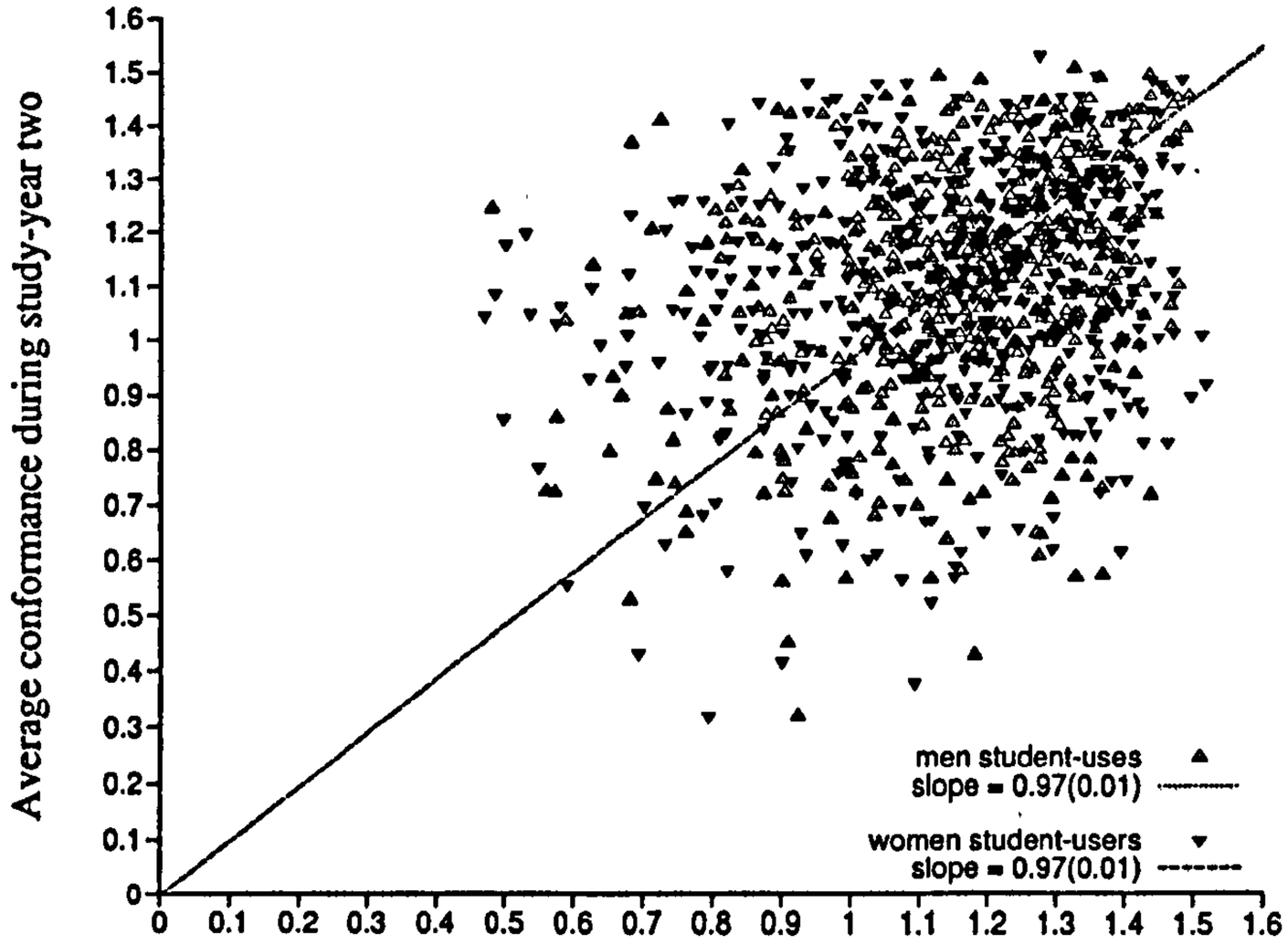

Figure E.6: Conditional distributions of student-user's average session-conformance by-gender

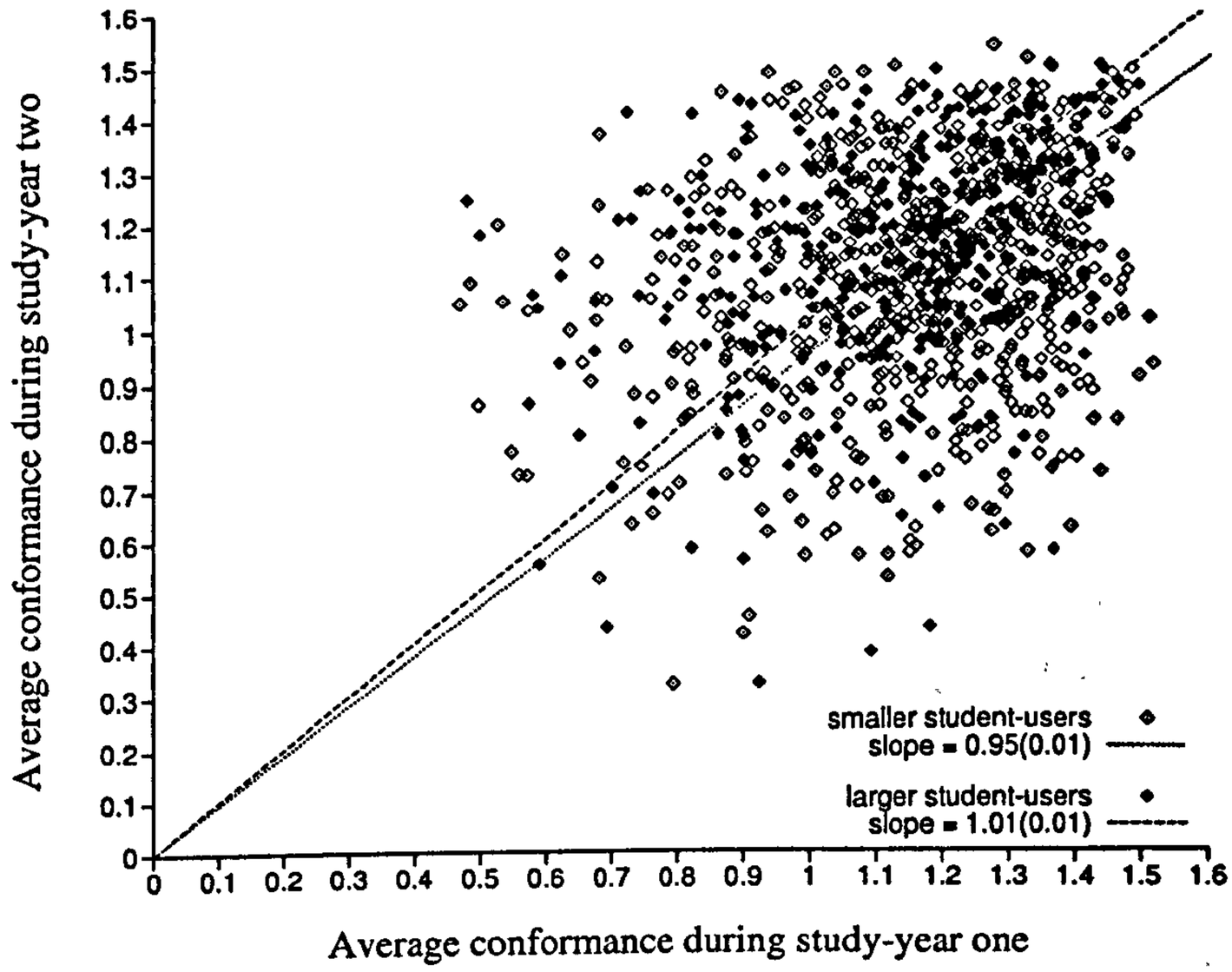

Figure E.7: Conditional distributions of student-user's average session-conformance by-joint-session-rate

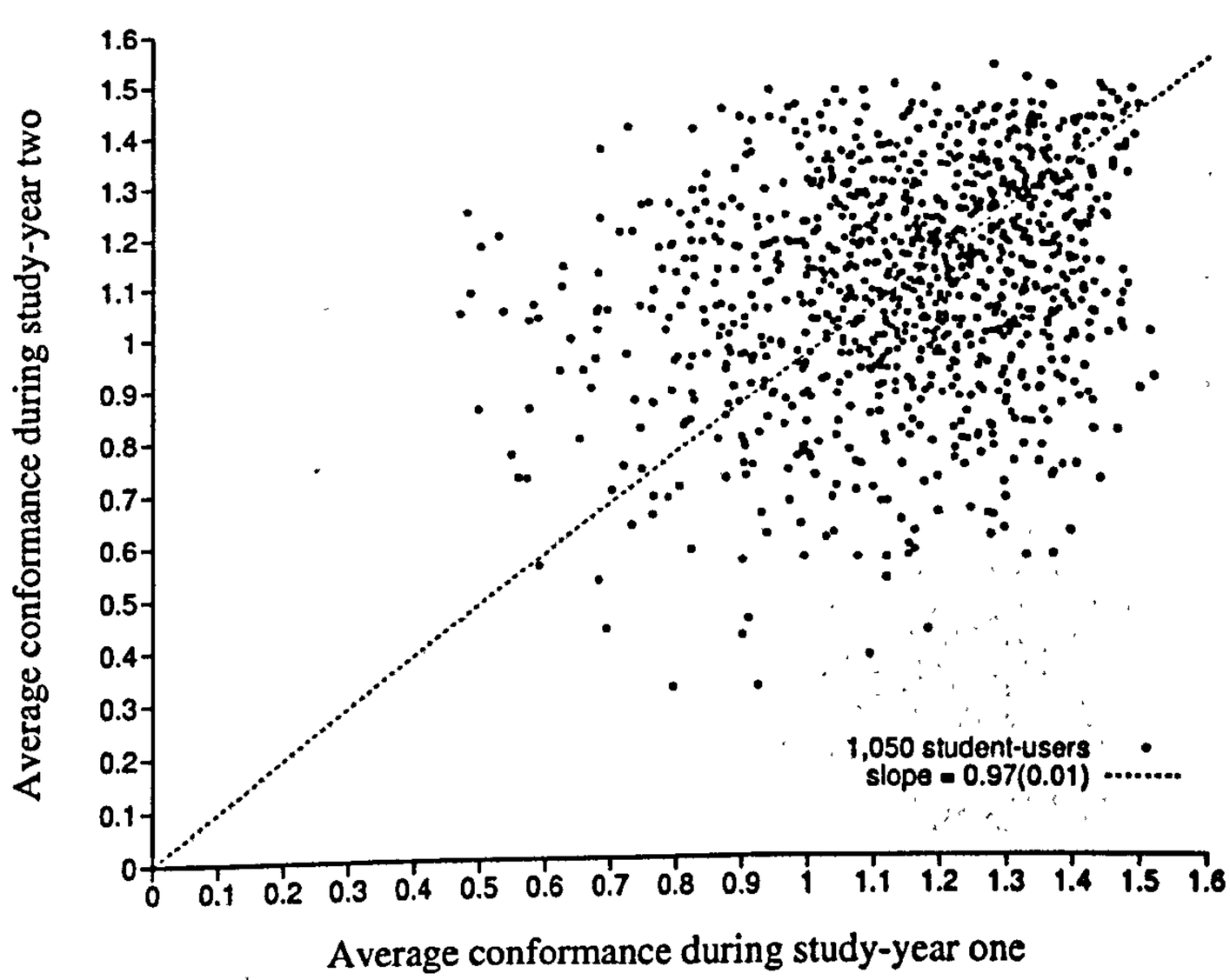

Figure E.8: Conditional distribution of student-user's average session-conformance

average query-click proportion

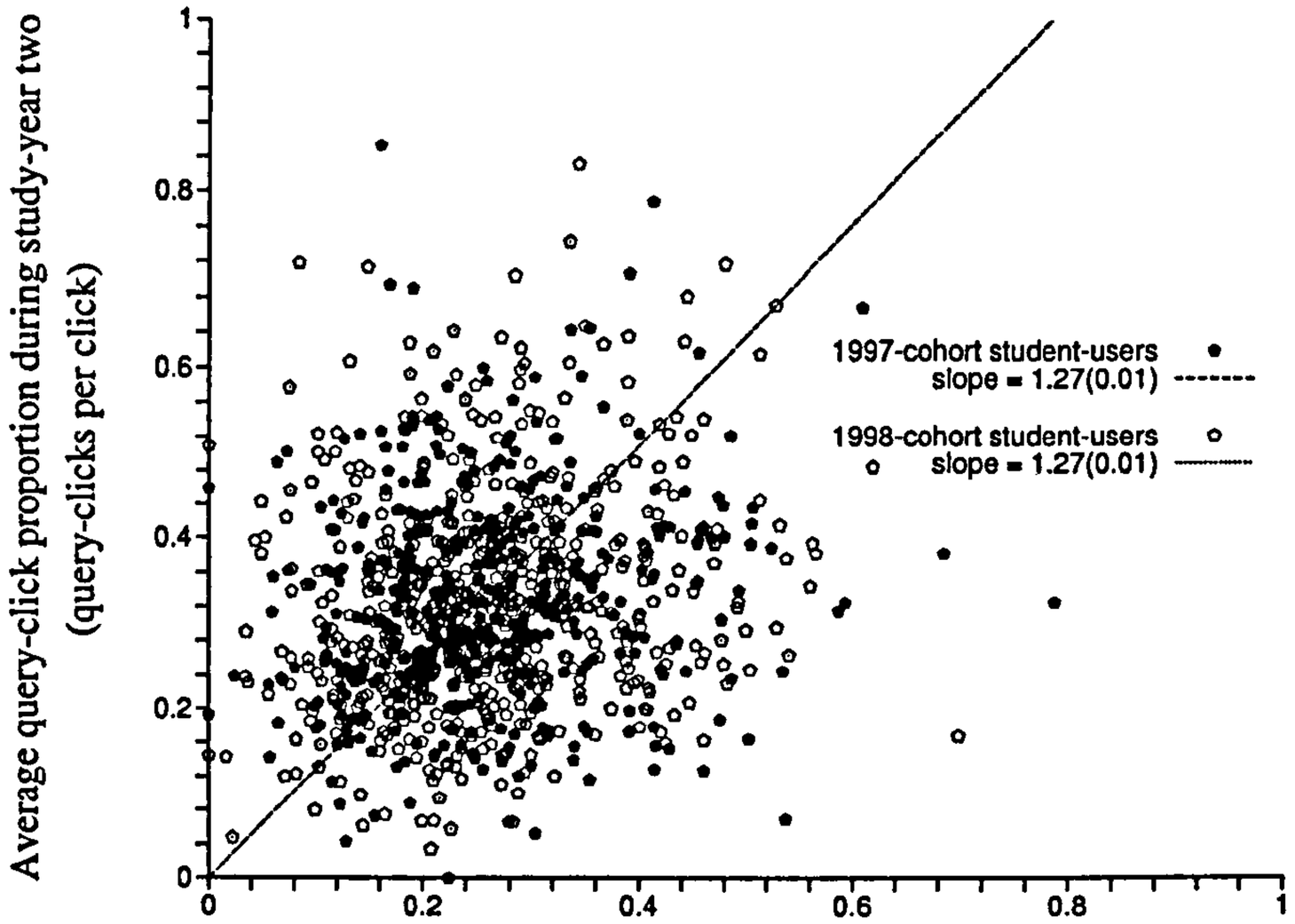

Figure E.9: Conditional distributions of student-user's average query-click proportion by-cohort

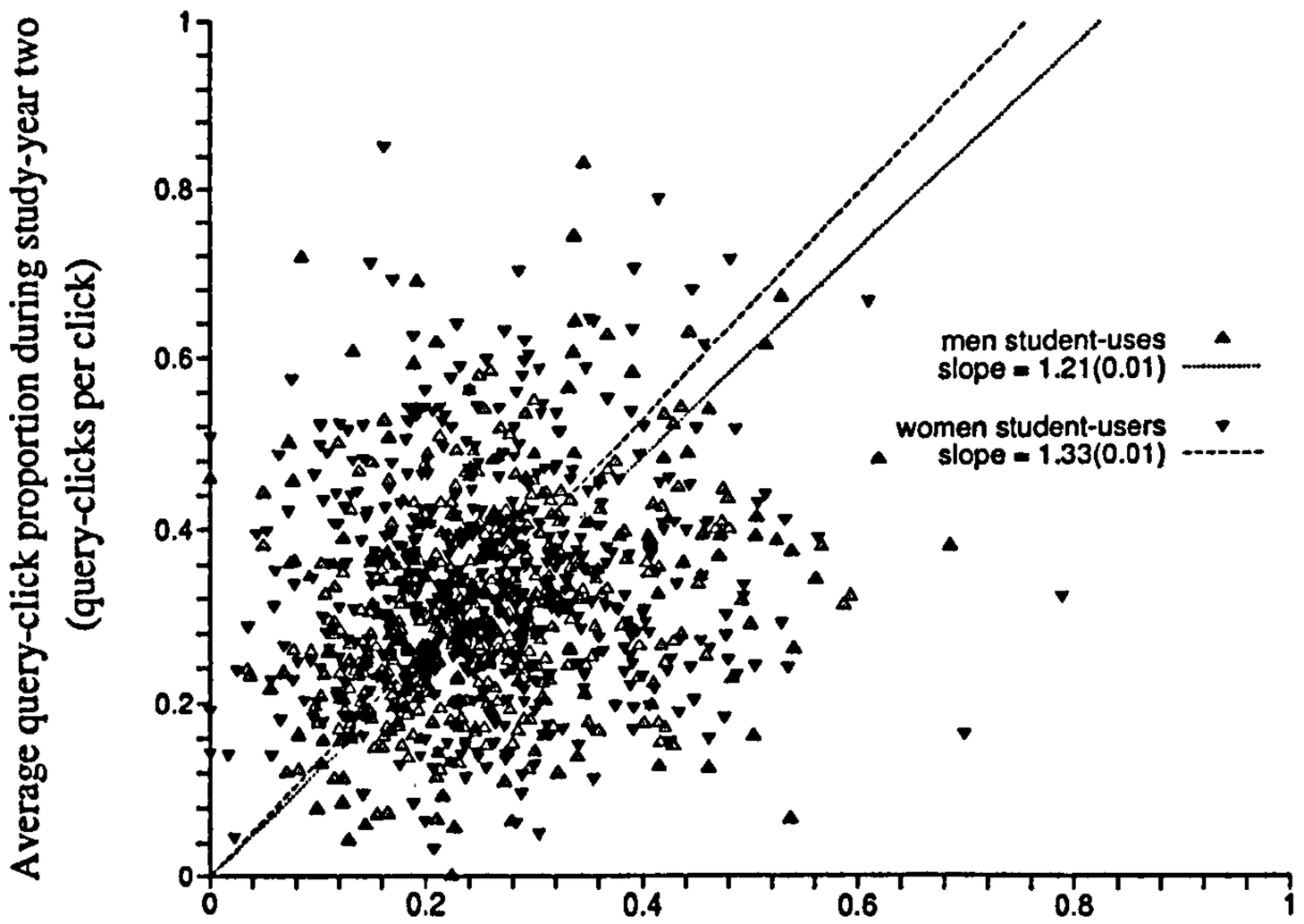

Figure E.10: Conditional distributions of student-user's average query-click proportion by-gender

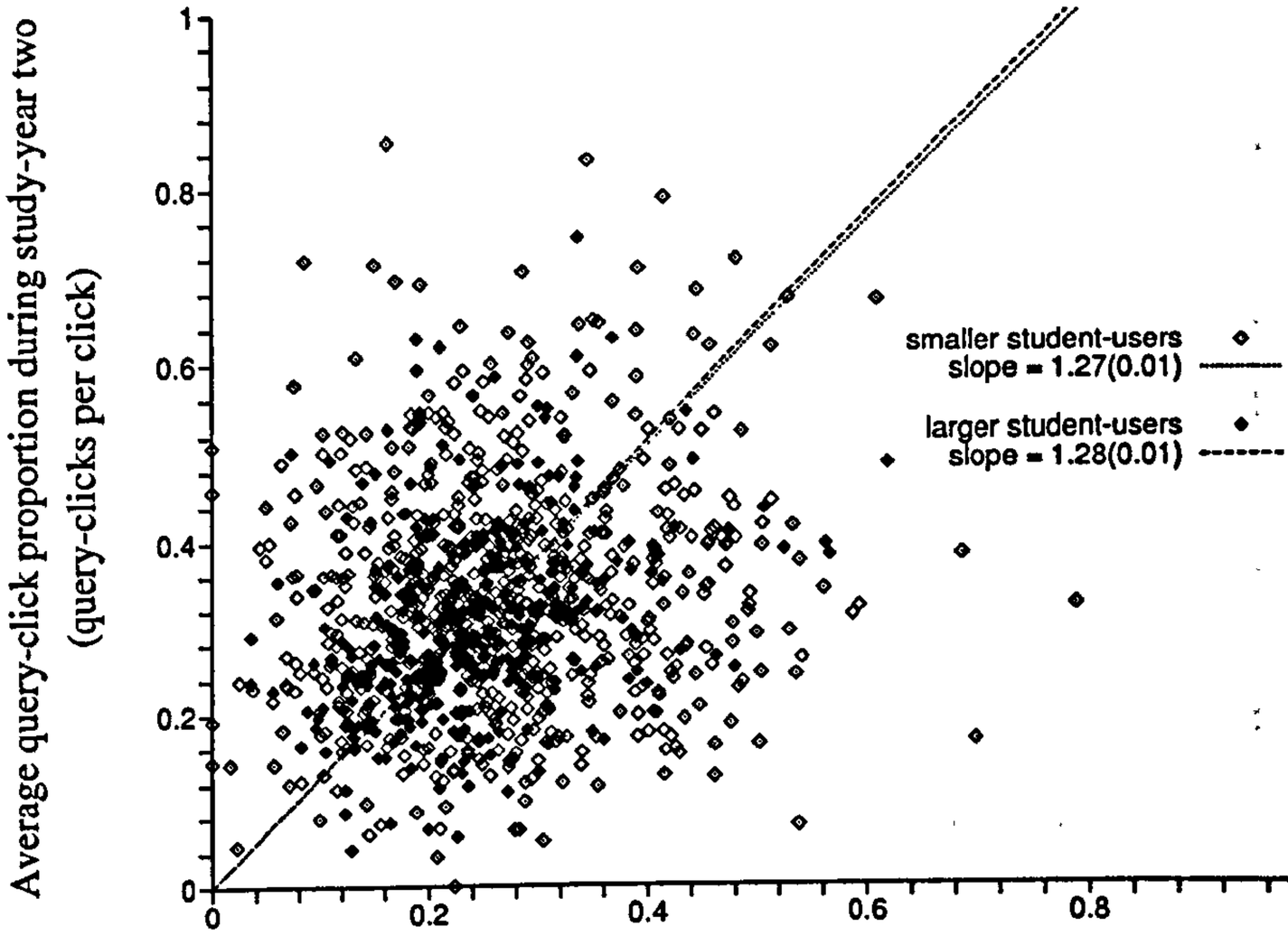

Figure E.11: Conditional distributions of student-user's average query-click proportion by-joint-session-rate

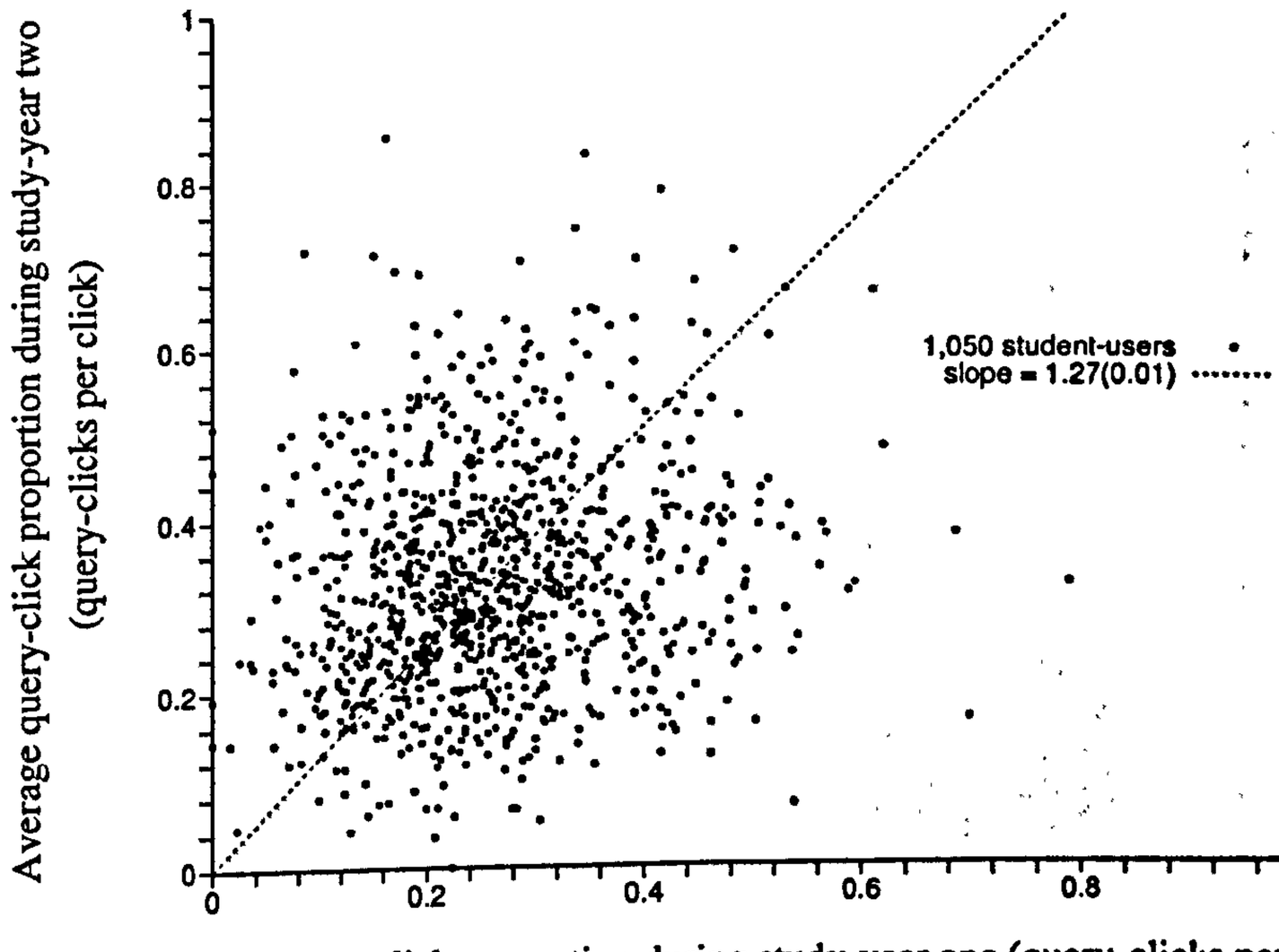

Figure E.12: Conditional distribution of student-user's average query-click proportion

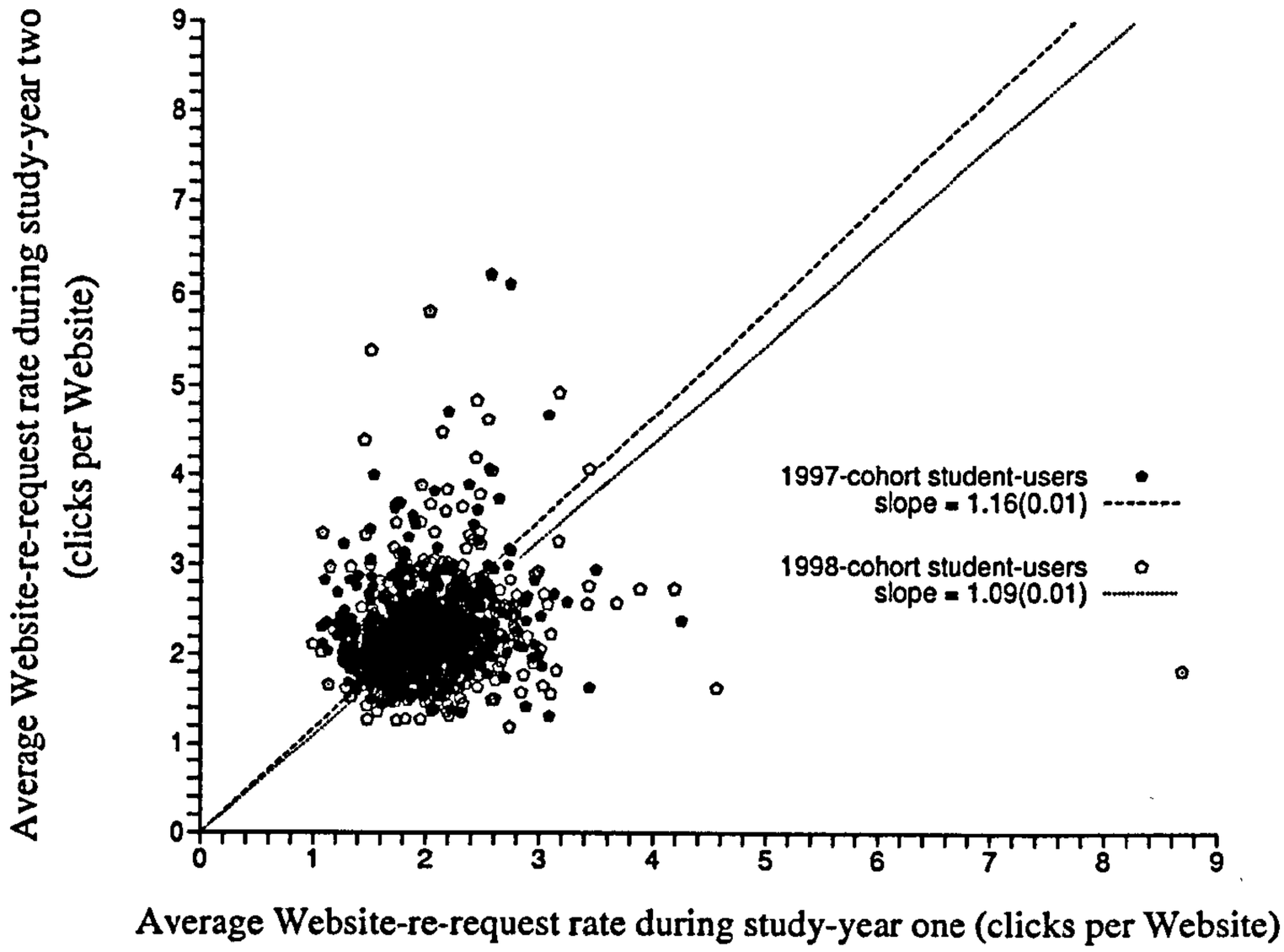

Figure E.13: Conditional distributions of student-user's average Website-re-request rate by-cohort

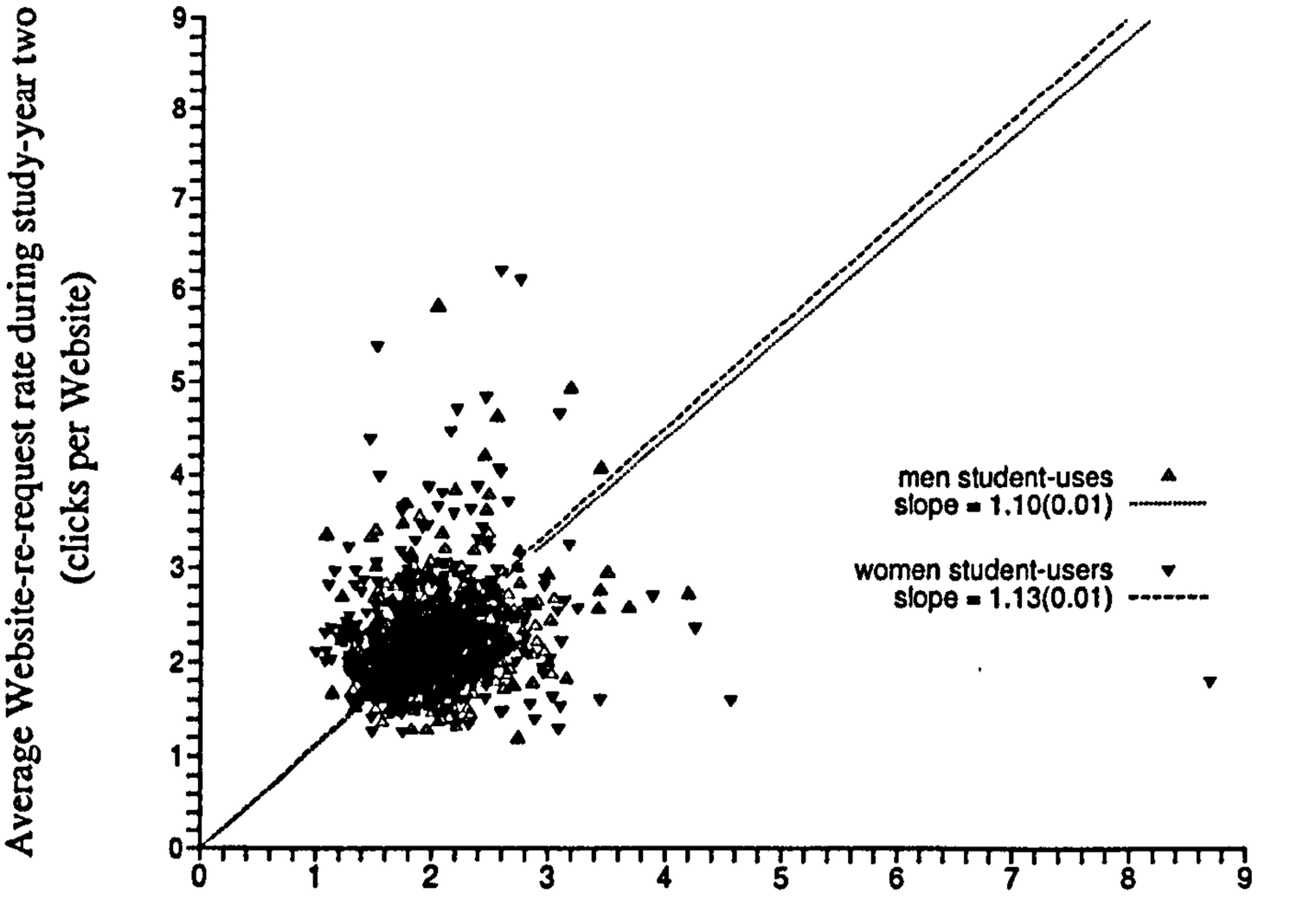

Figure E.14: Conditional distributions of student-user's average Website-re-request rates by-gender

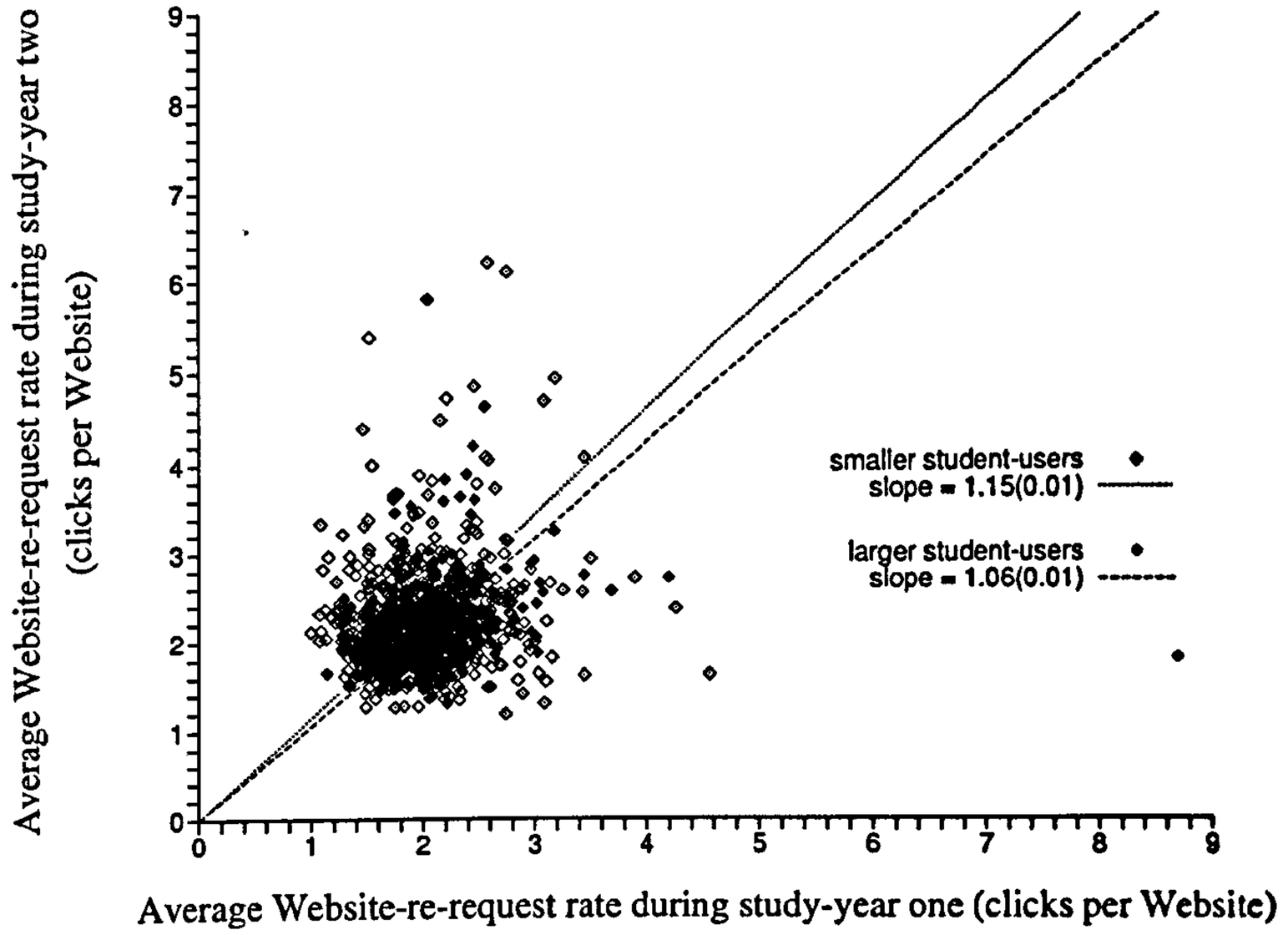

Figure E.15: Conditional distributions of student-user's average Website-re-request rates by-joint-session-rate

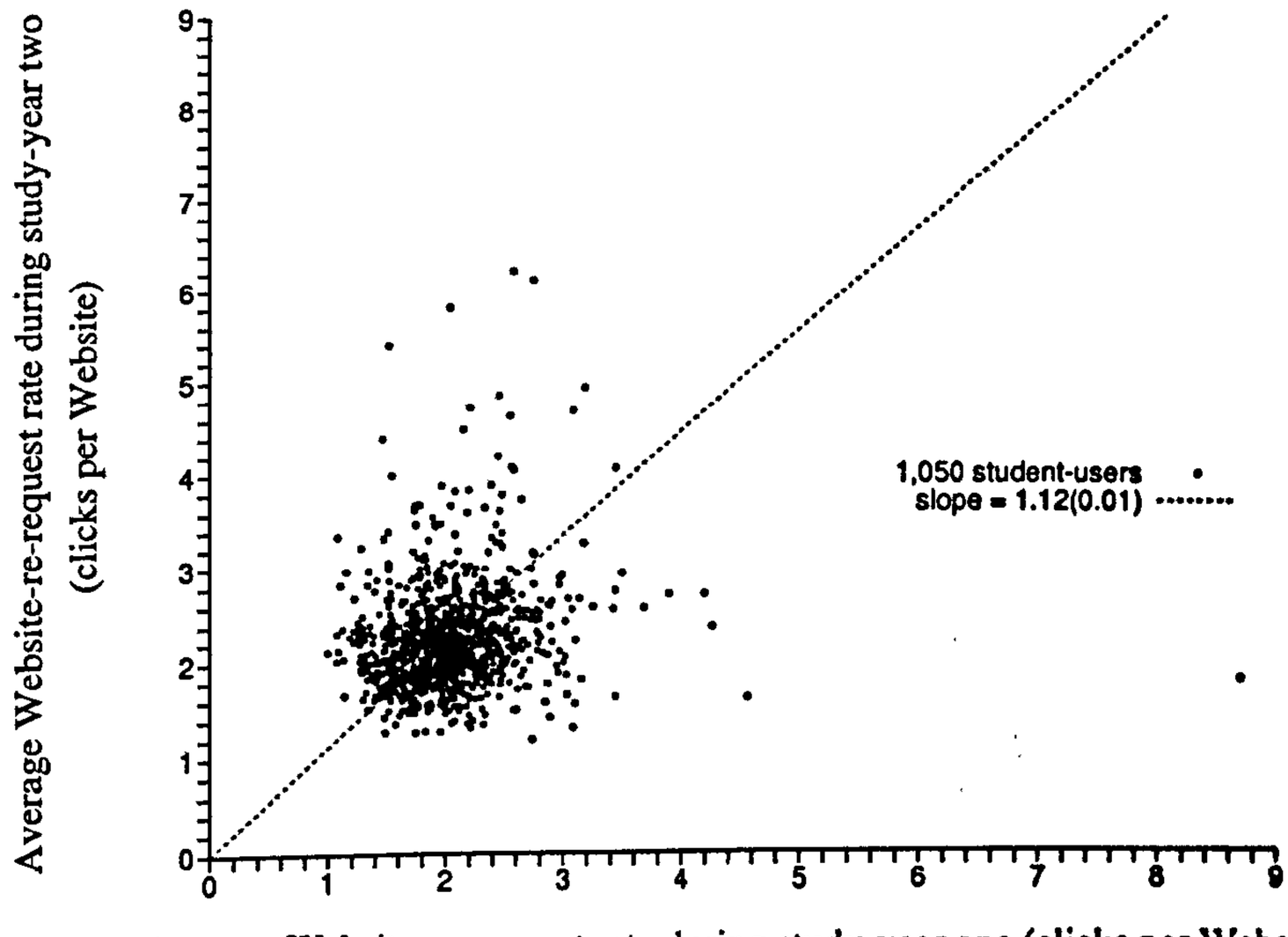

Figure E.16: Conditional distribution of student-user's average Website-re-request rate

average Webhost-persistence

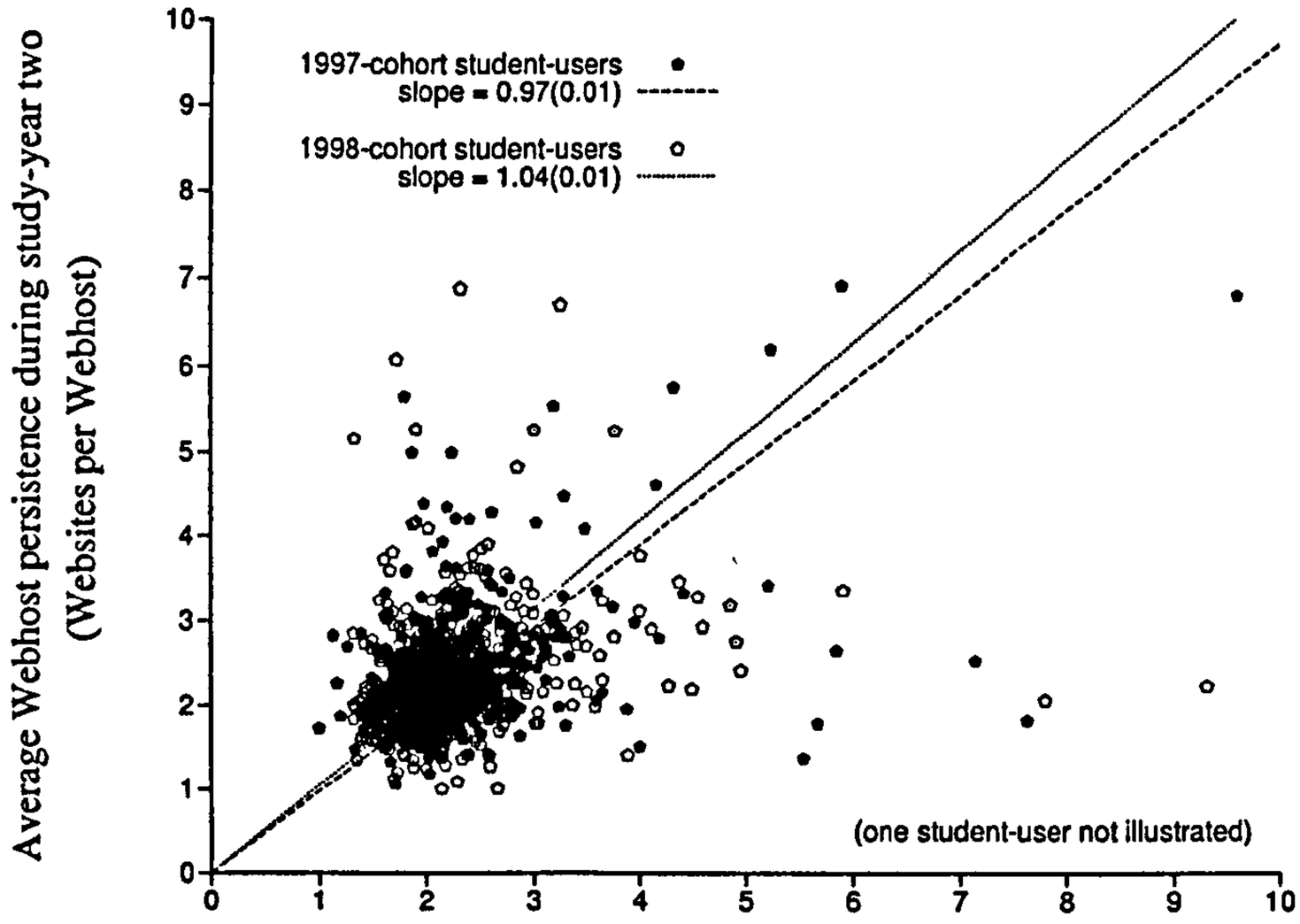

Figure E.17: Conditional distributions of student-user's average Webhost-persistence by-cohort

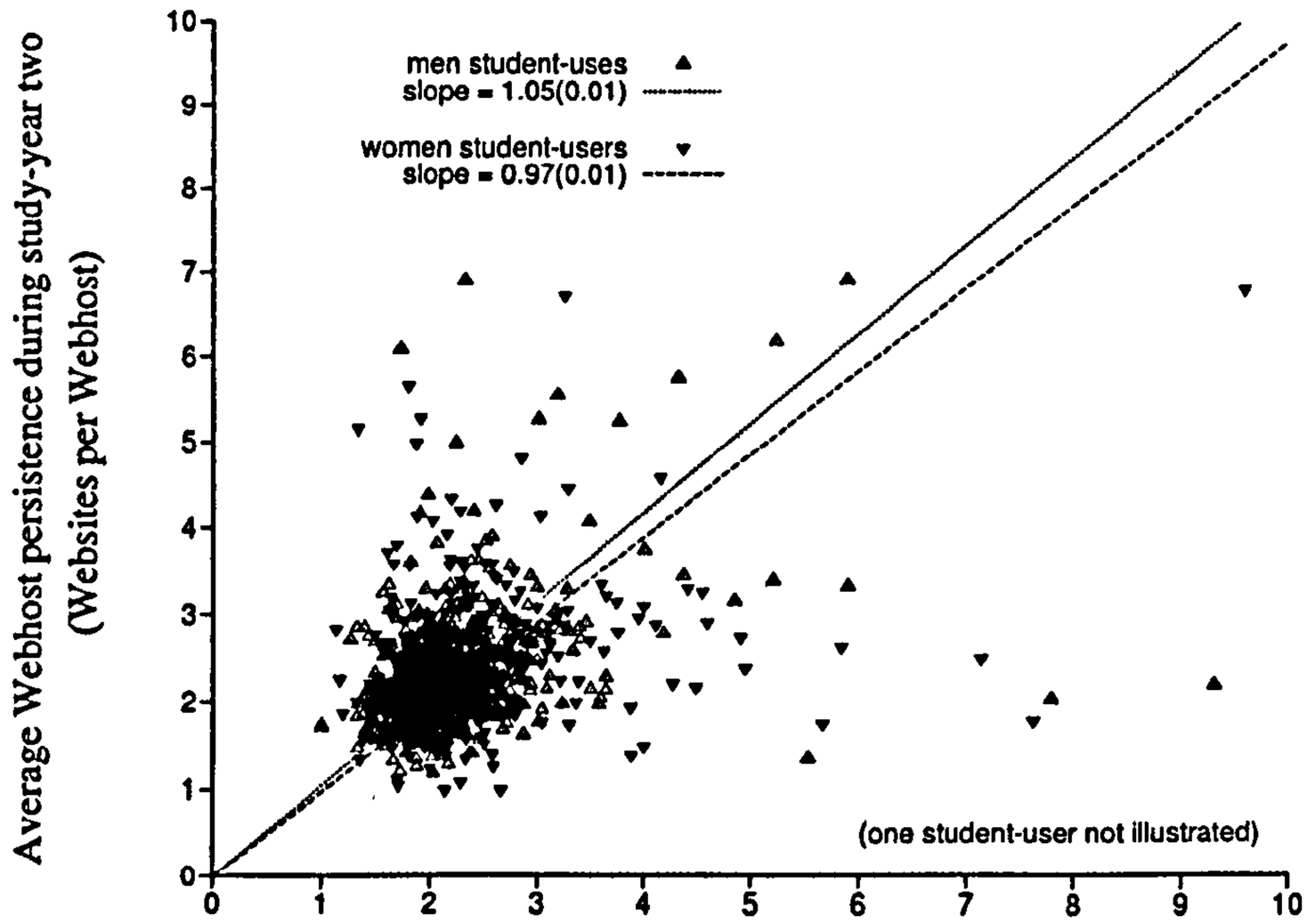

Figure E.18: Conditional distributions of student-user's average Webhost-persistence by-gender

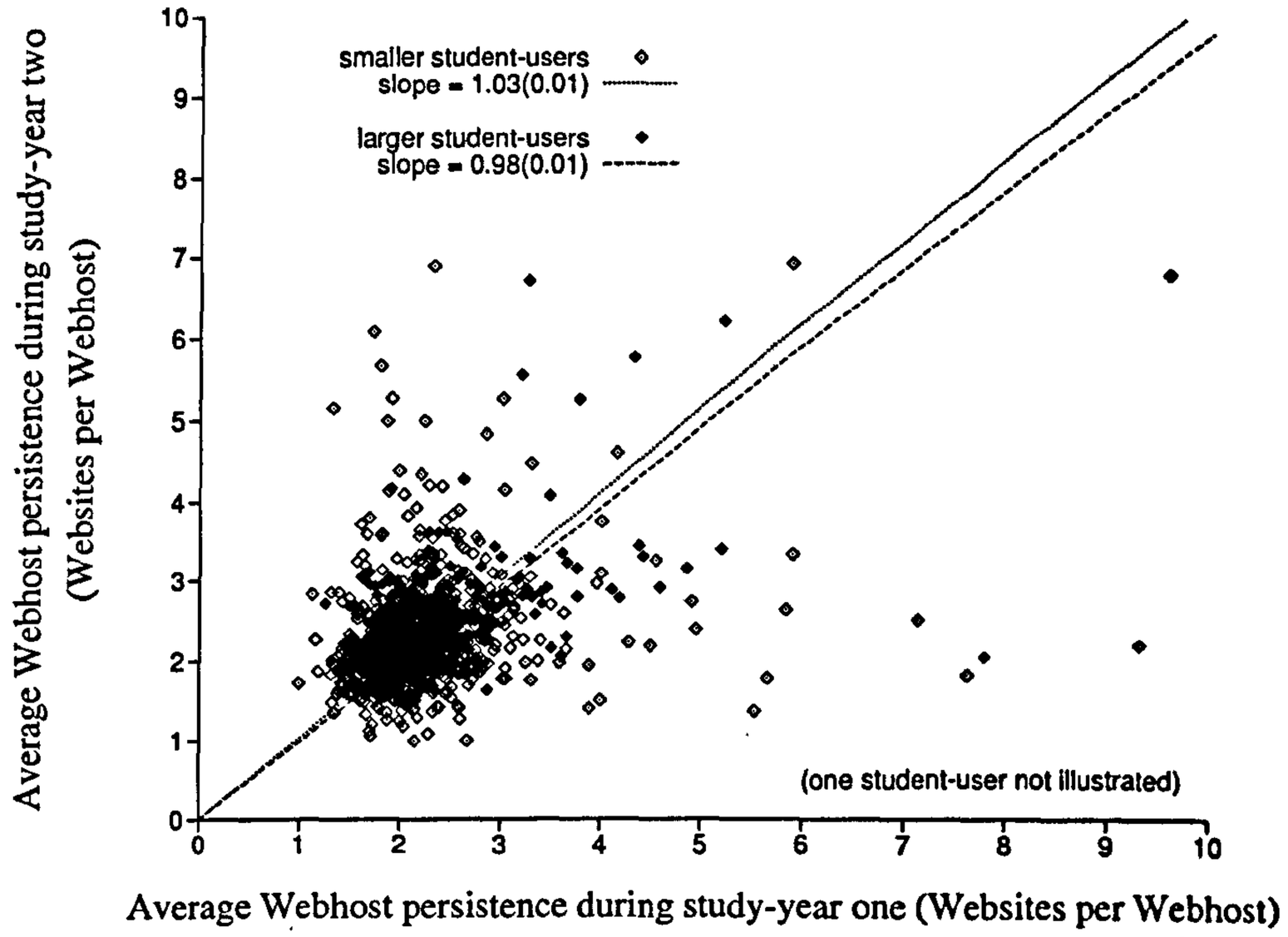

Figure E.19: Conditional distributions of student-user's average Webhost-persistence by-joint-session-rate

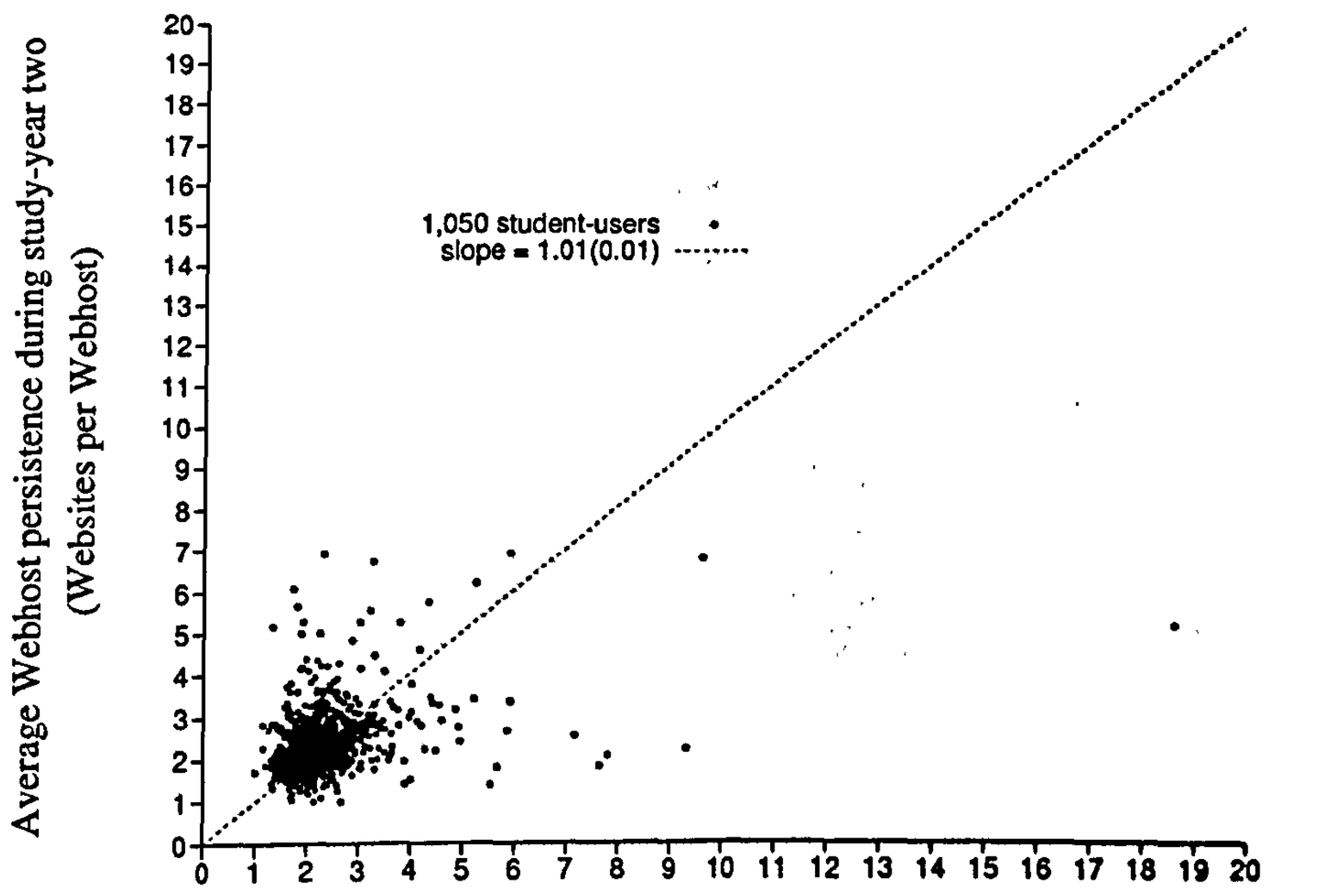

Figure E.20: Conditional distribution of student-user's average Webhost-persistence

Website-trajectory slope

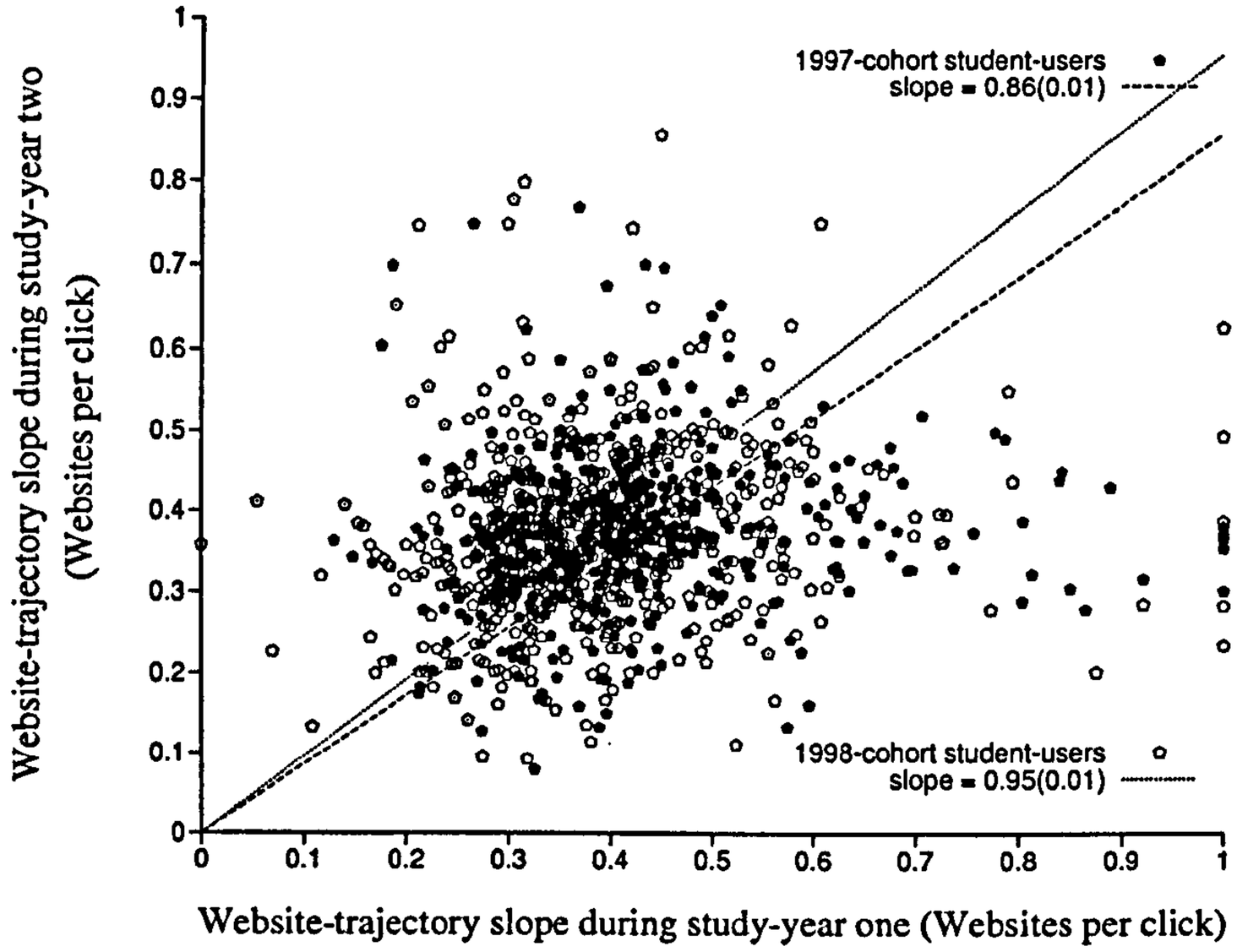

Figure E.21: Conditional distributions of student-user's Website-trajectory slope by-cohort

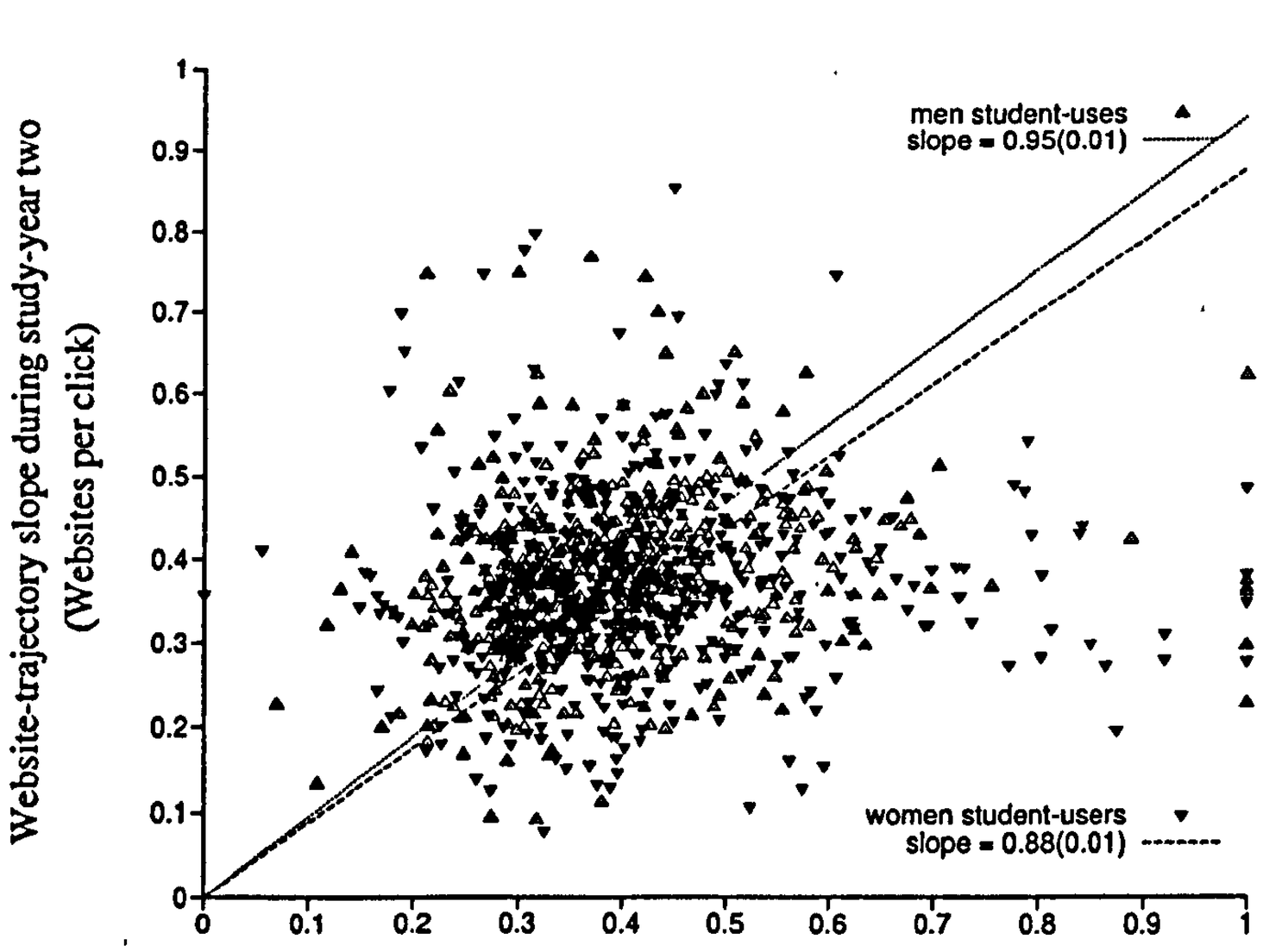

Figure E.22: Conditional distributions of student-user's Website-trajectory slope by-gender

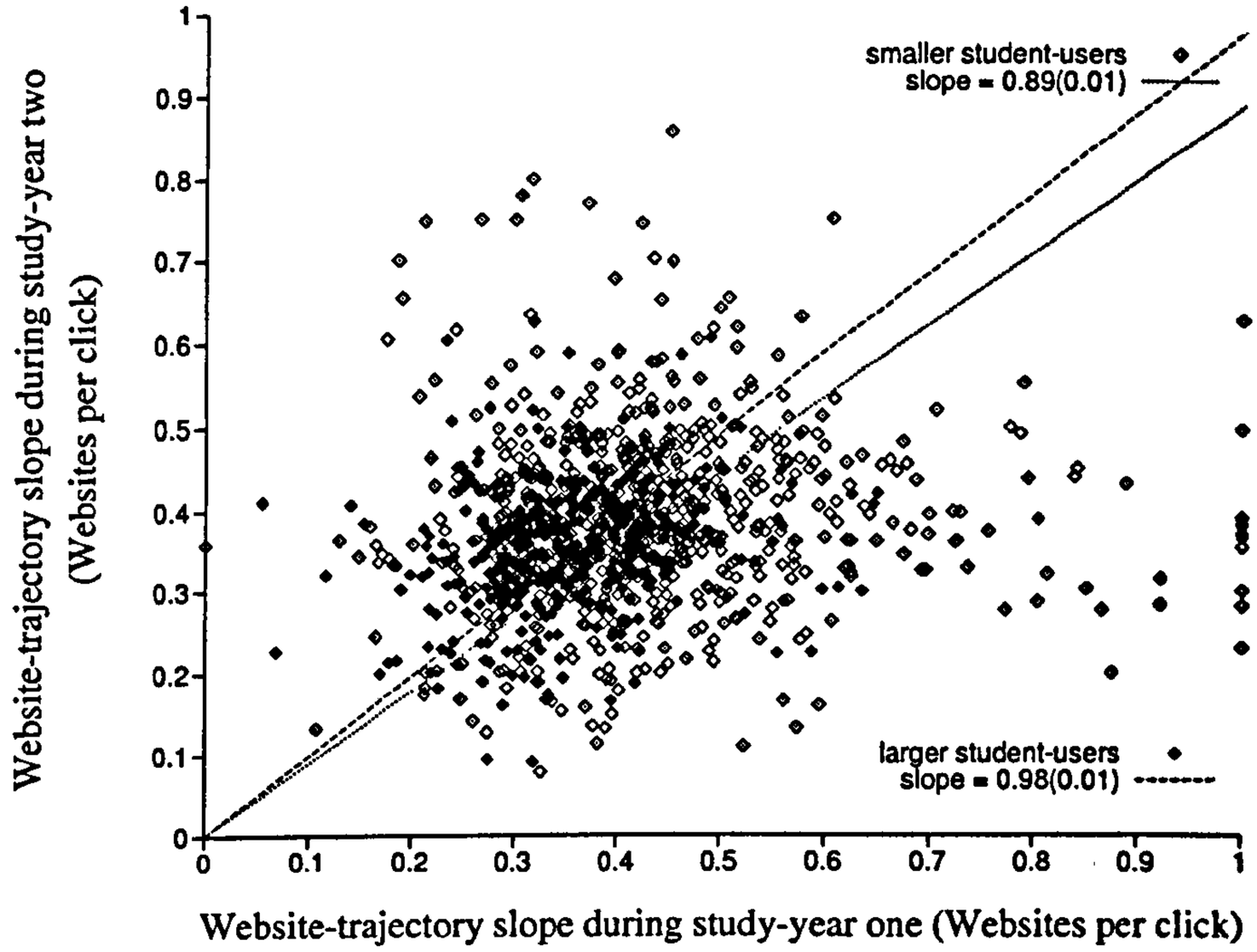

Figure E.23: Conditional distributions of student-user's Website-trajectory slope by-joint-session-rate

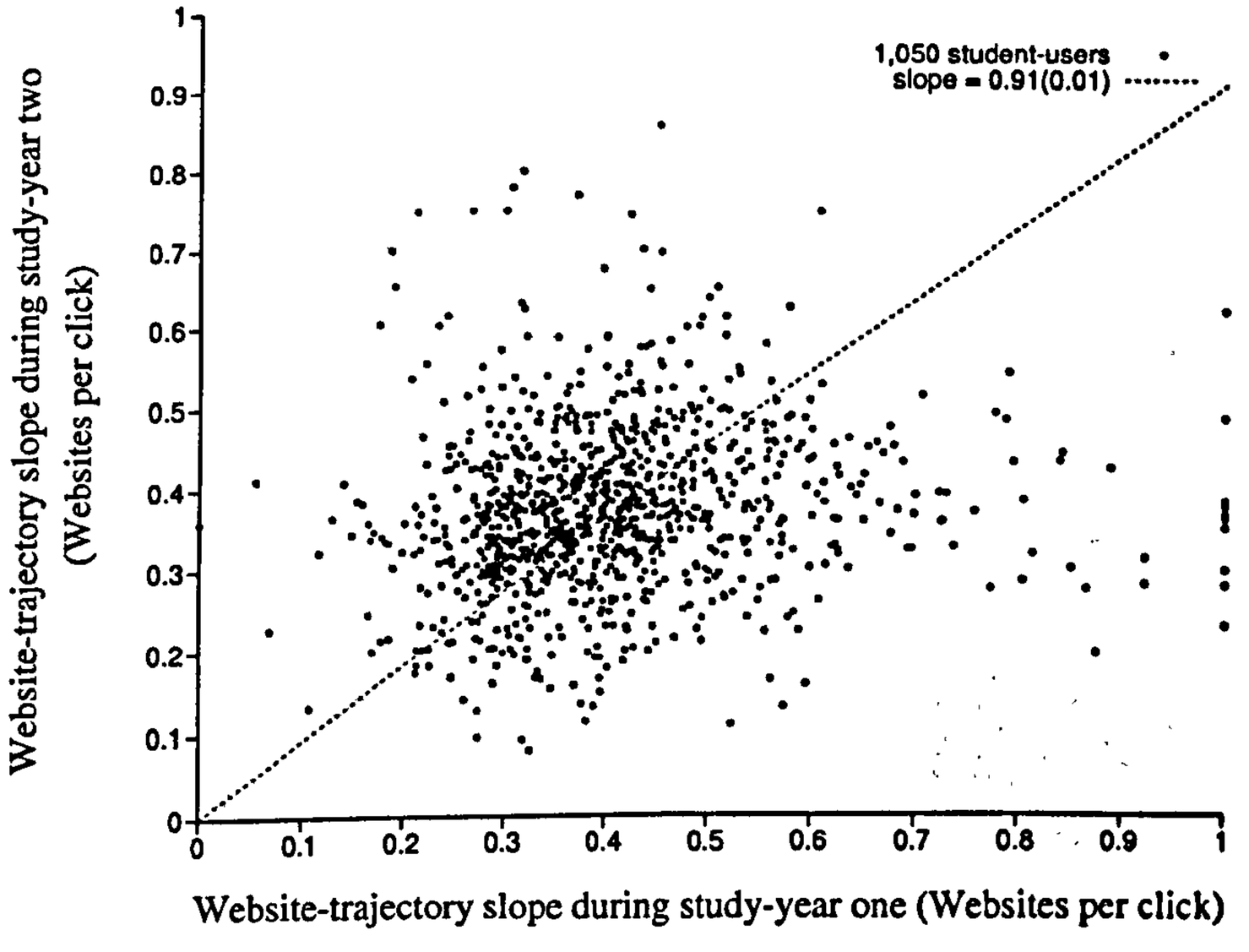

Figure E.24: Conditional distribution of student-user's Website-trajectory slope

average search-query proportion

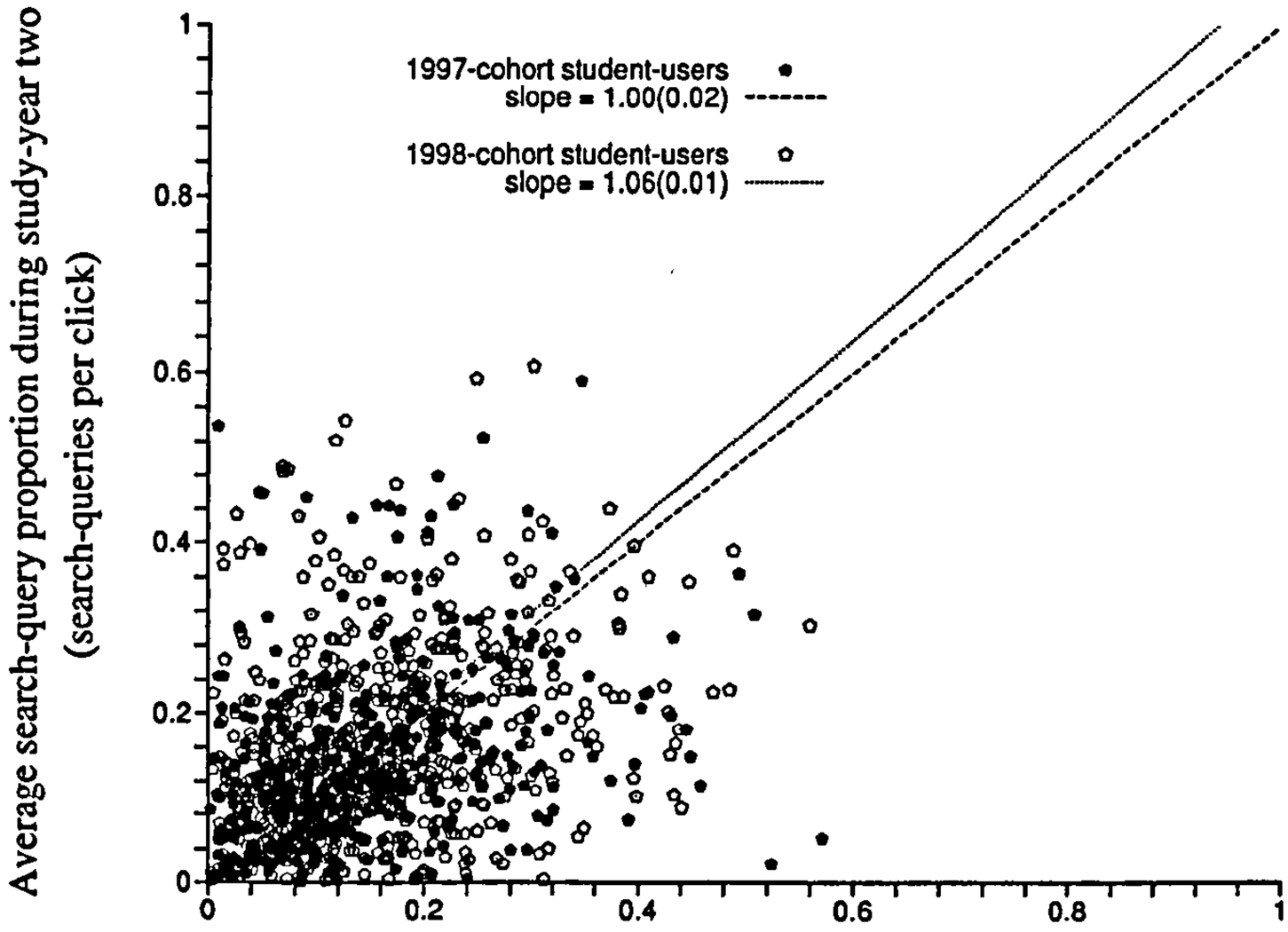

Figure E.25: Conditional distributions of search-user's average search-query proportion by-cohort

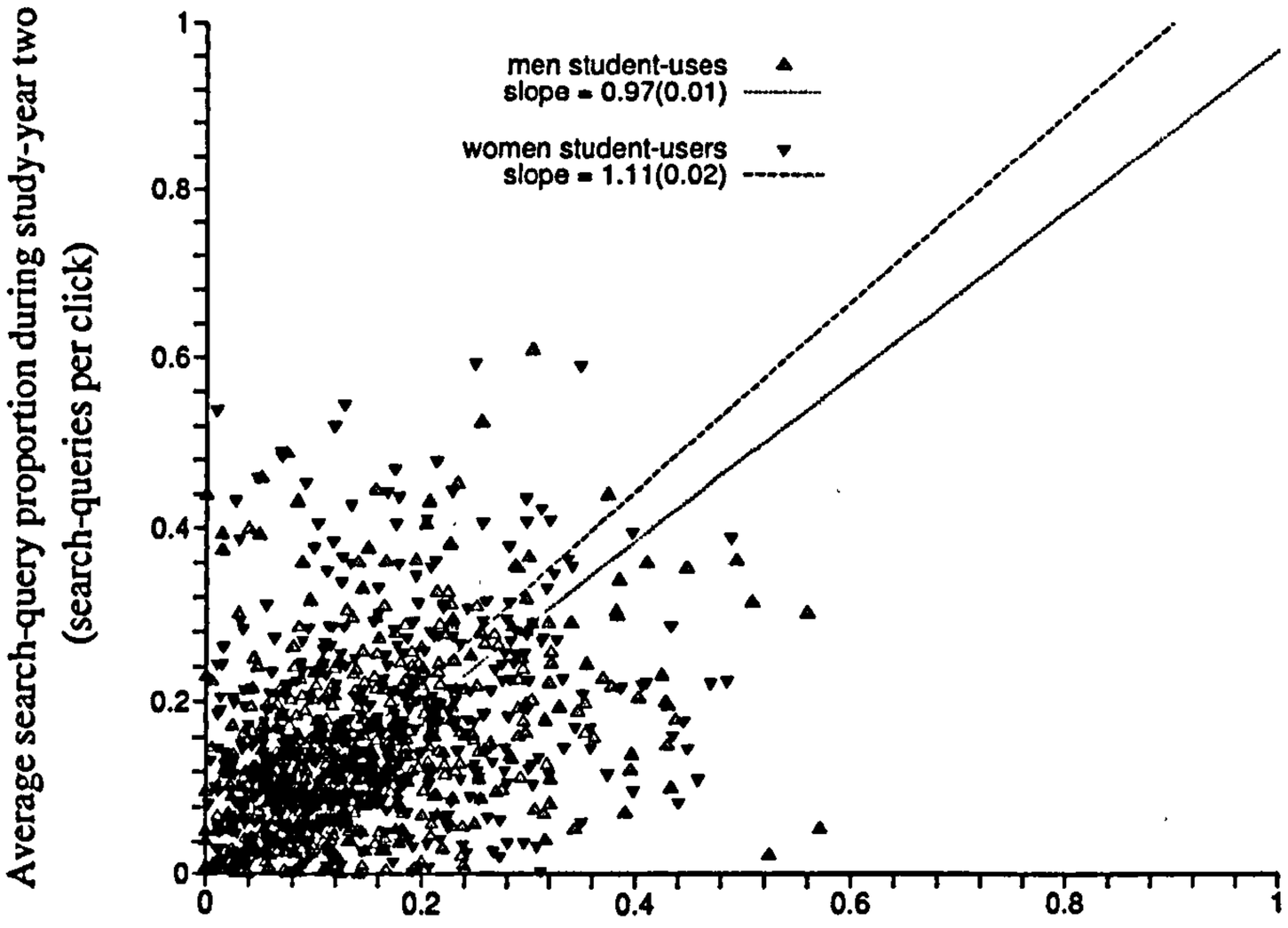

Figure E.26: Conditional distributions of search-user's search-query proportion by-gender

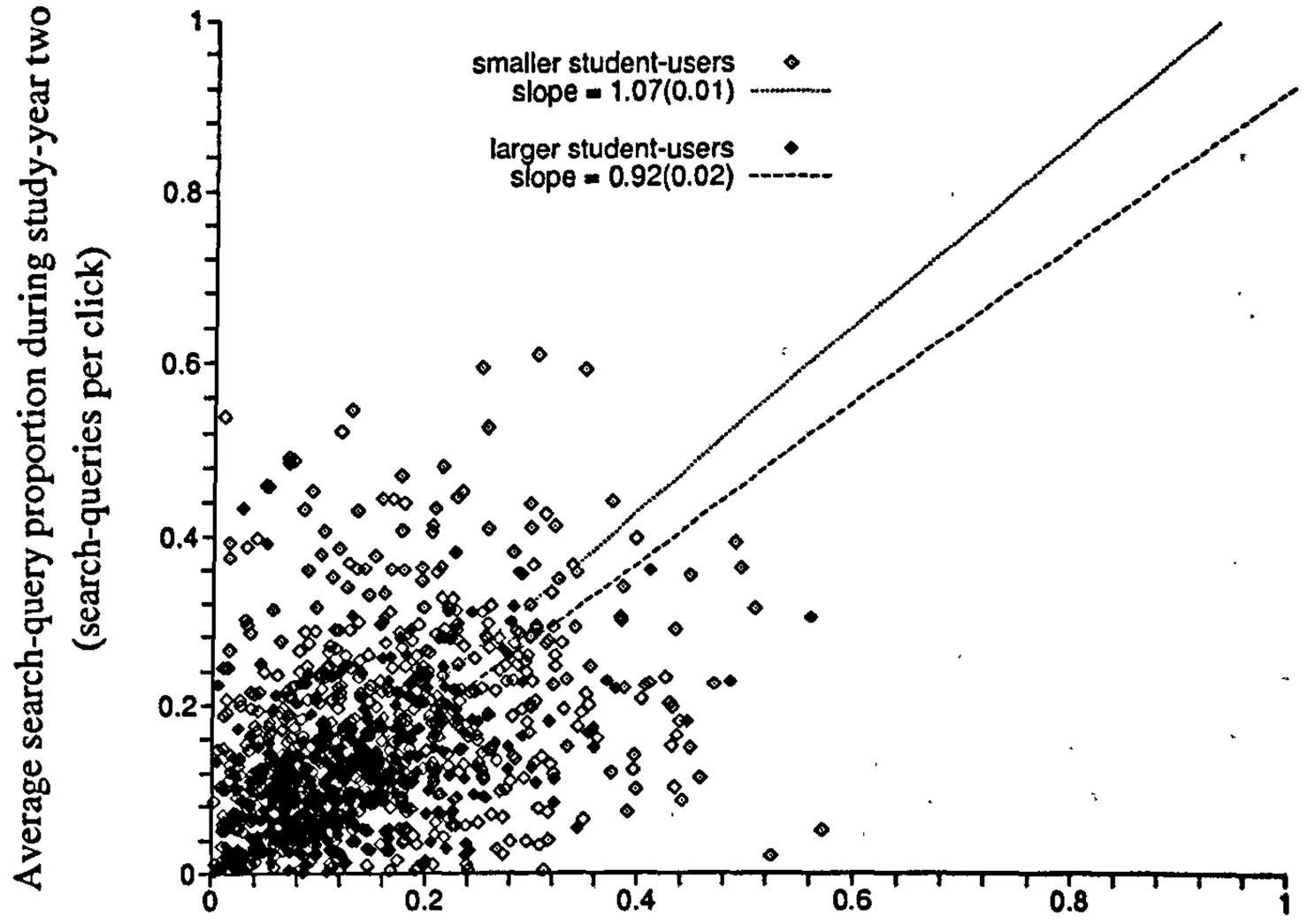

Figure E.27: Conditional distributions of search-user's average search-query proportion by-joint-session-rate

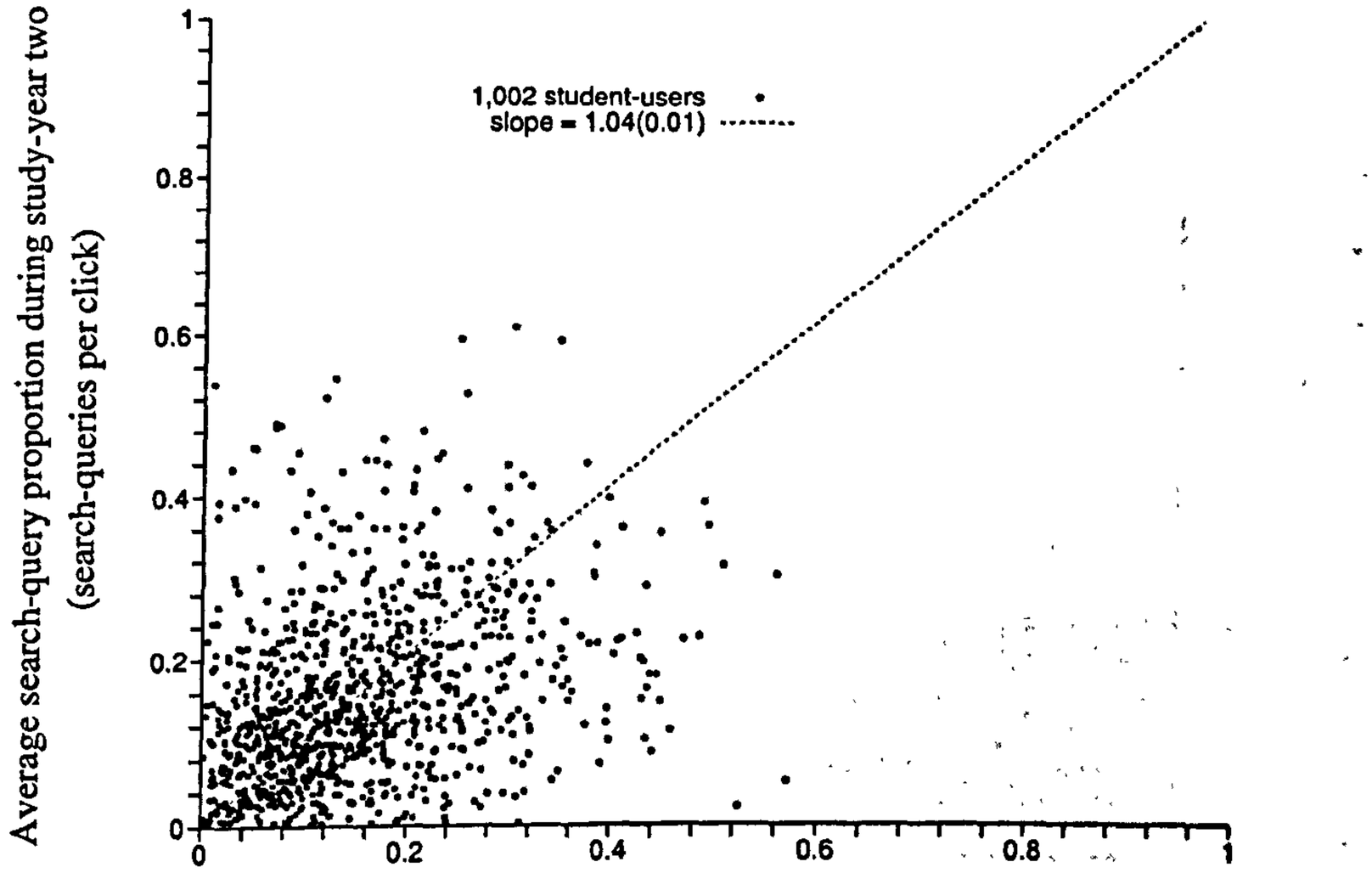

Figure E.28: Conditional distribution of search-user's average search-query proportion

search-session proportion

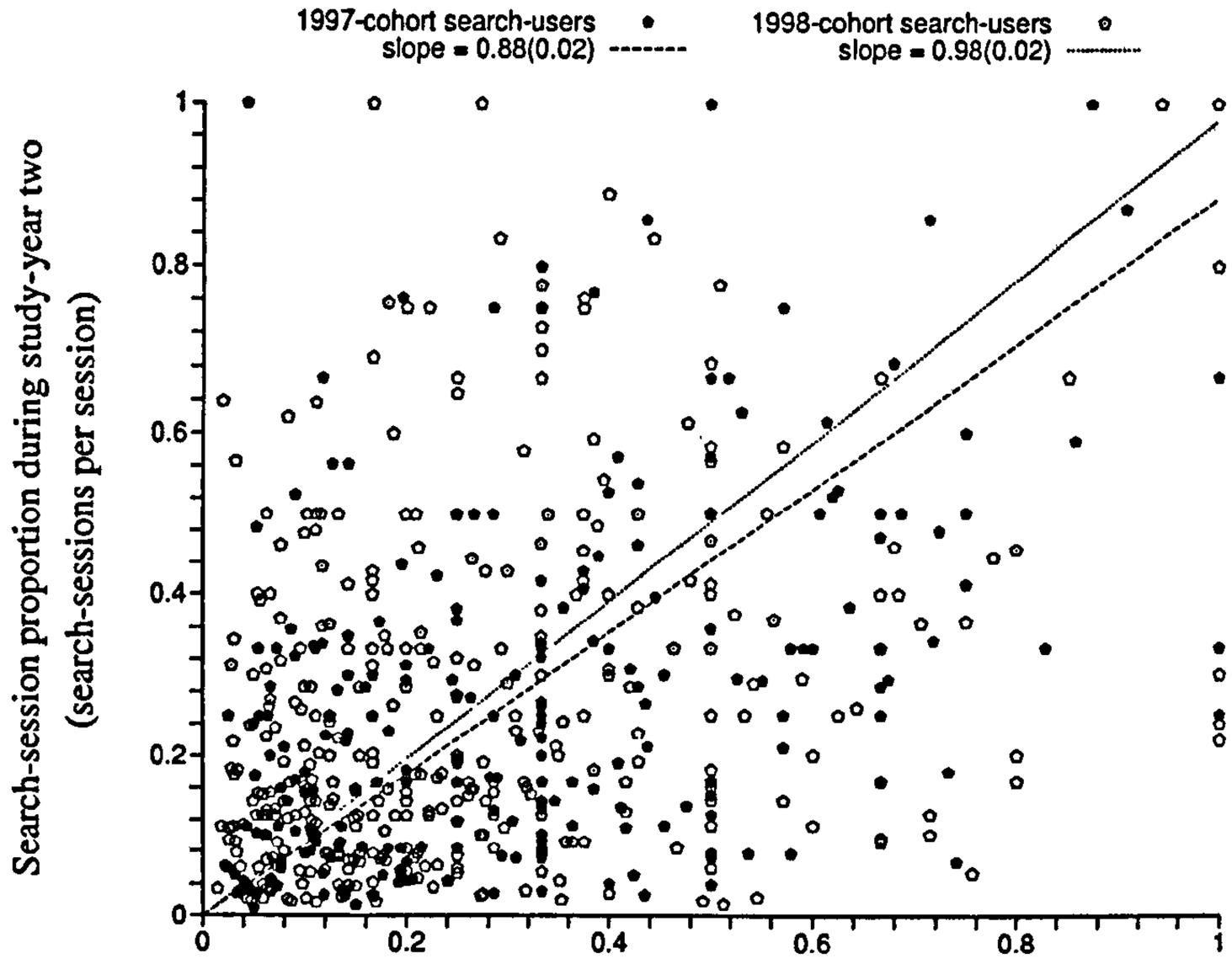

Figure E.29: Conditional distributions of search-user's search-session proportion by-cohort

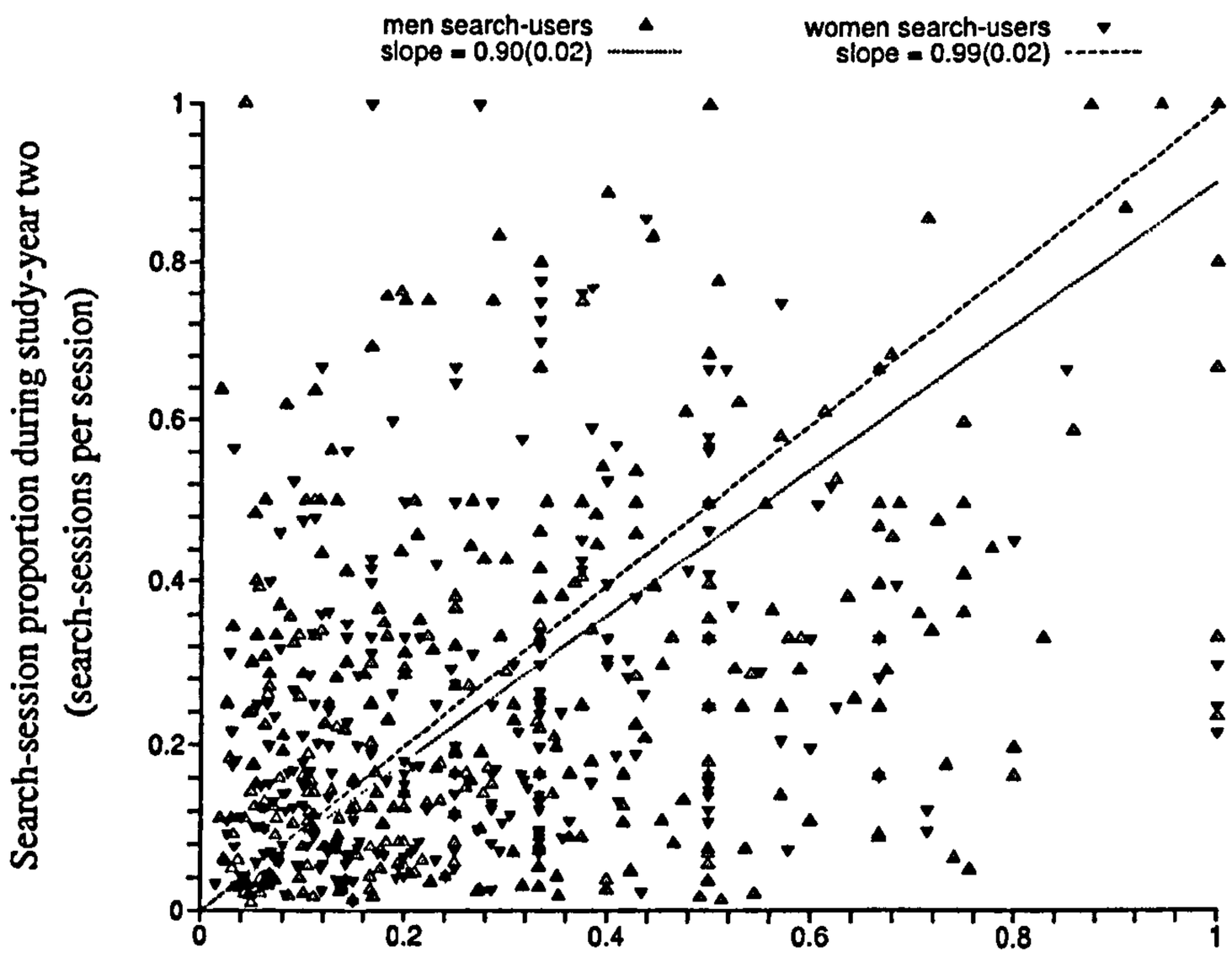

Figure E.30: Conditional distributions of search-user's search-session proportion by-gender

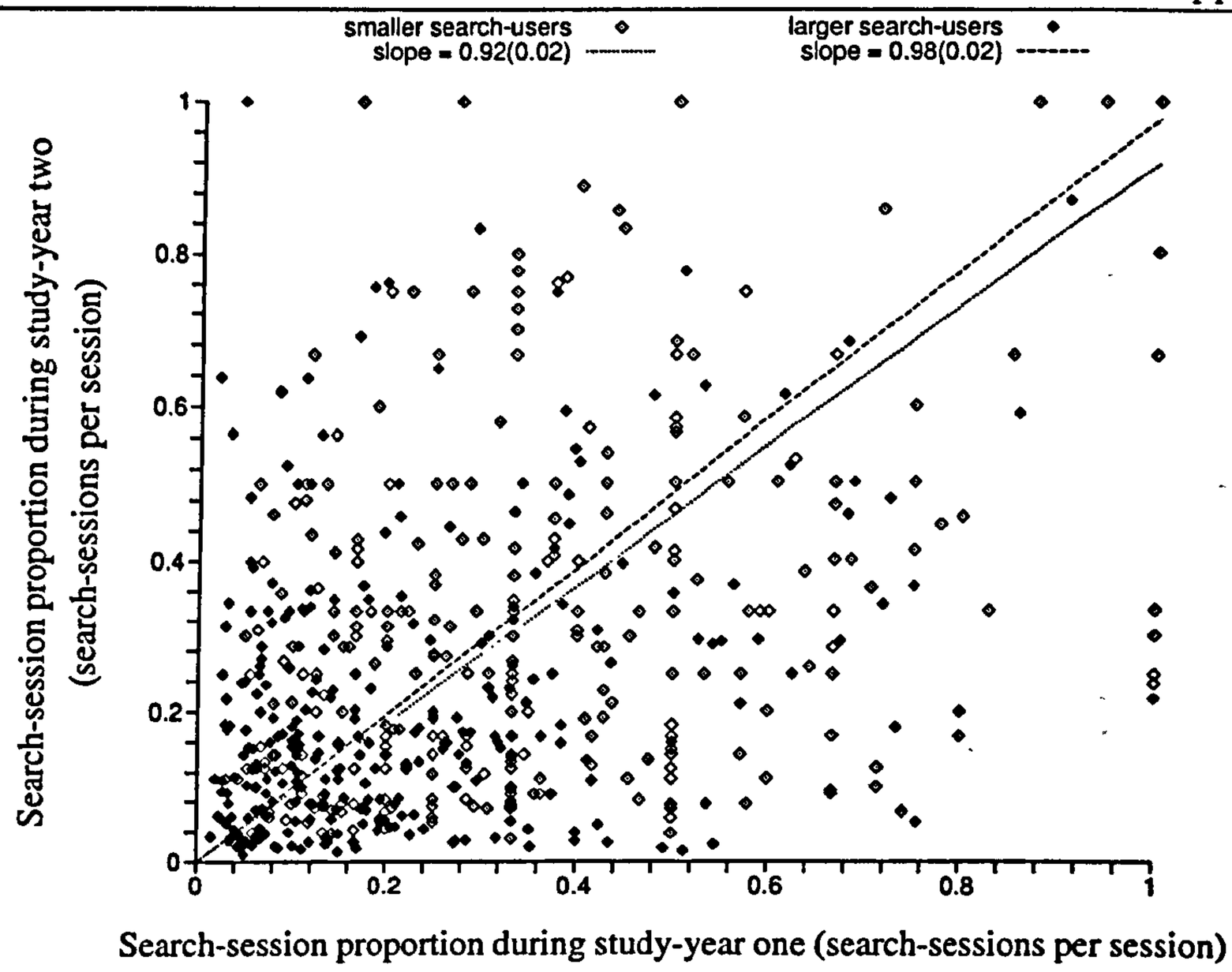

Figure E.31: Conditional distributions of search-user's search-session proportion by-joint-session-rate

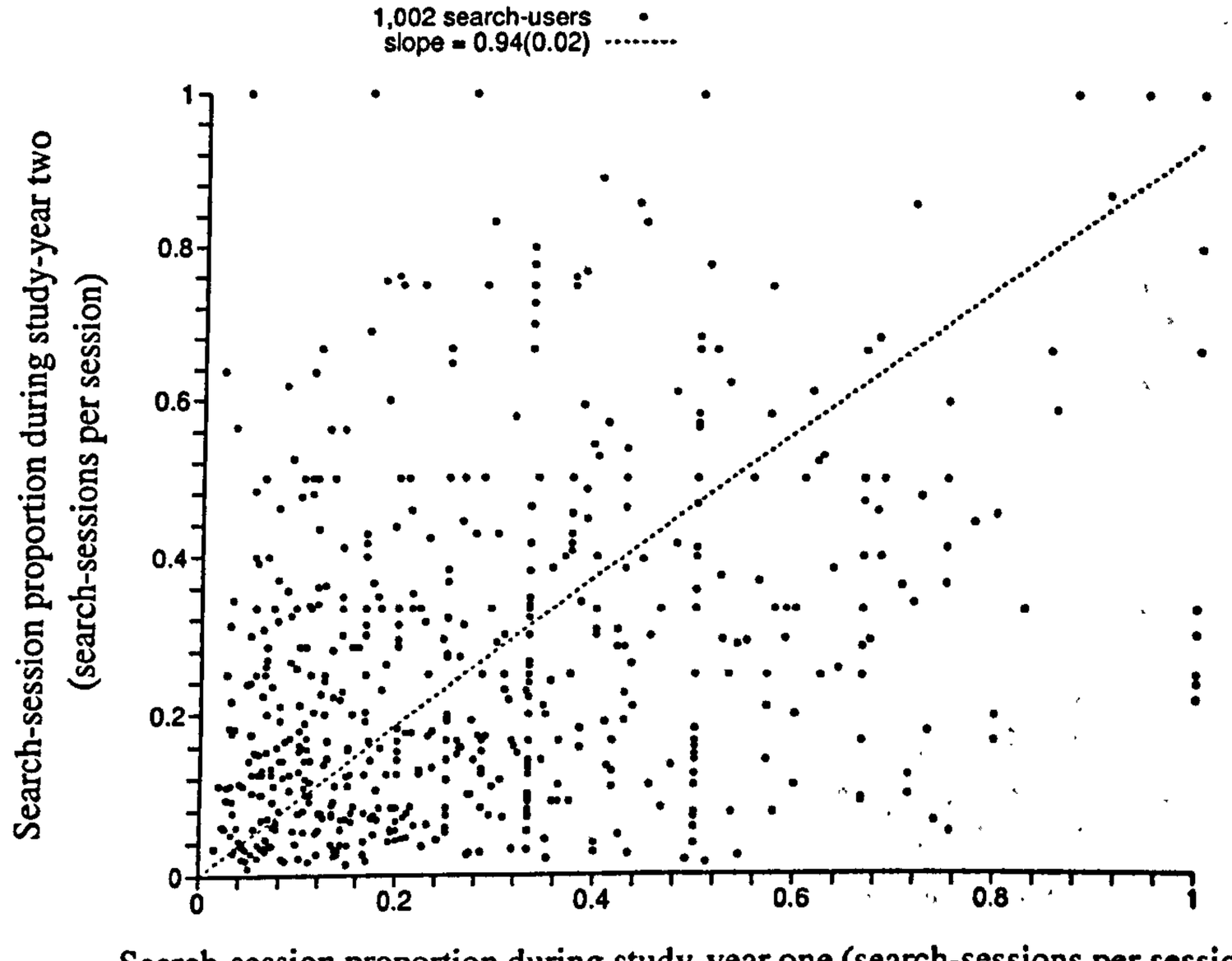

Figure E.32: Conditional distribution of search-user's search-session proportion

average search-query count

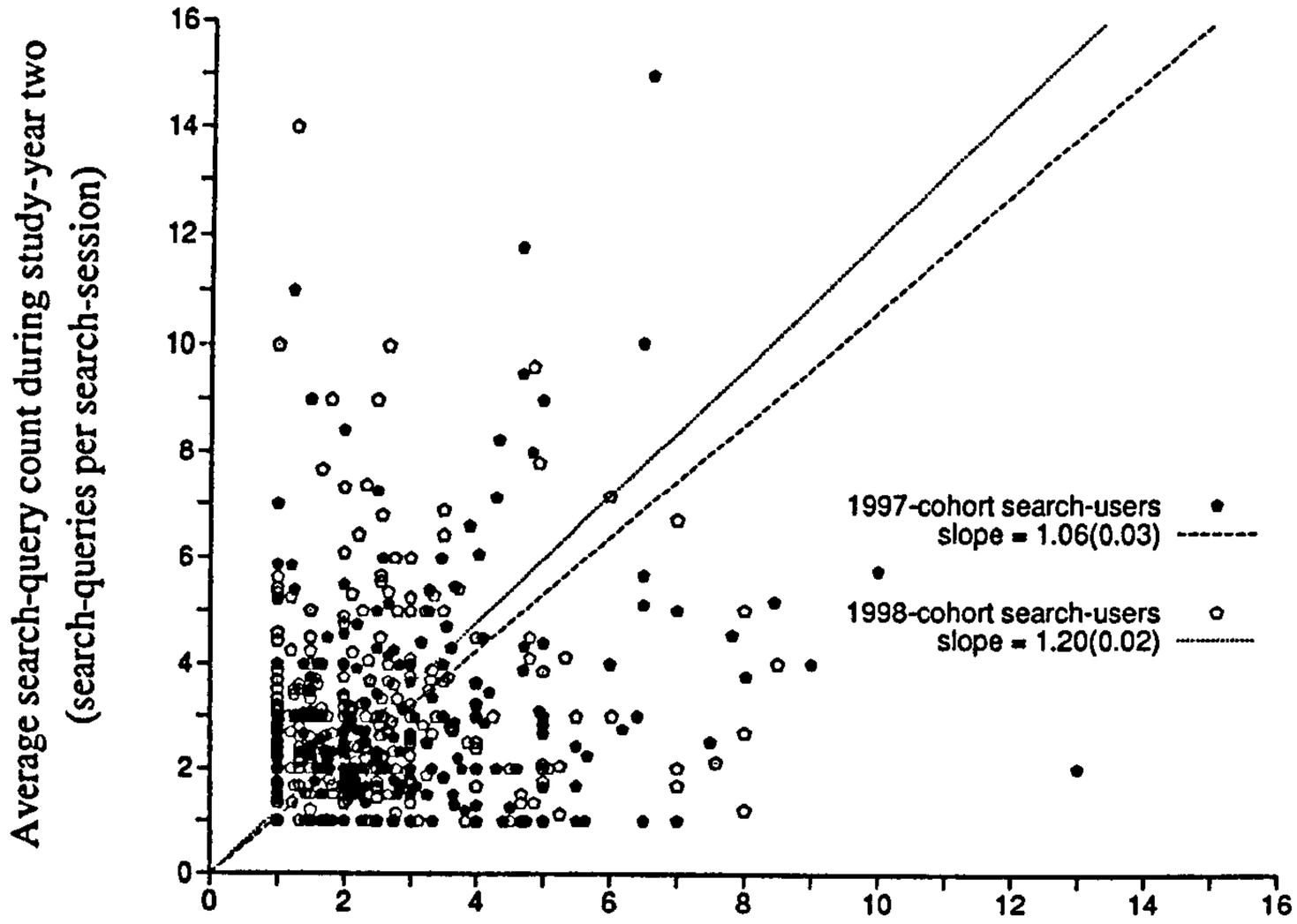

Figure E.33: Conditional distributions of AltaVista-Excite sample search-user's average search-query count by-cohort

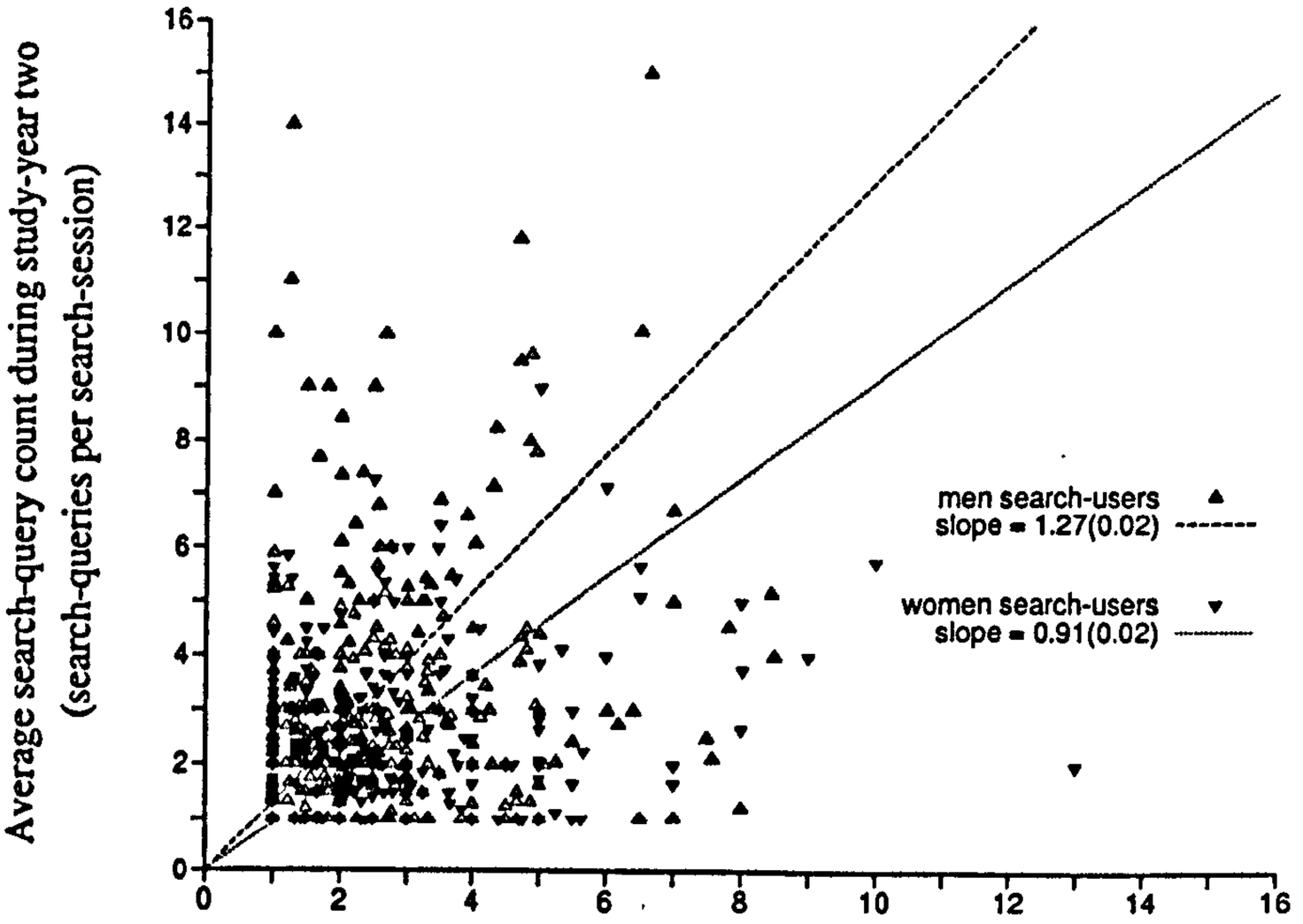

Figure E.34: Conditional distributions of AltaVista-Excite sample search-user's average search-query count by-gender

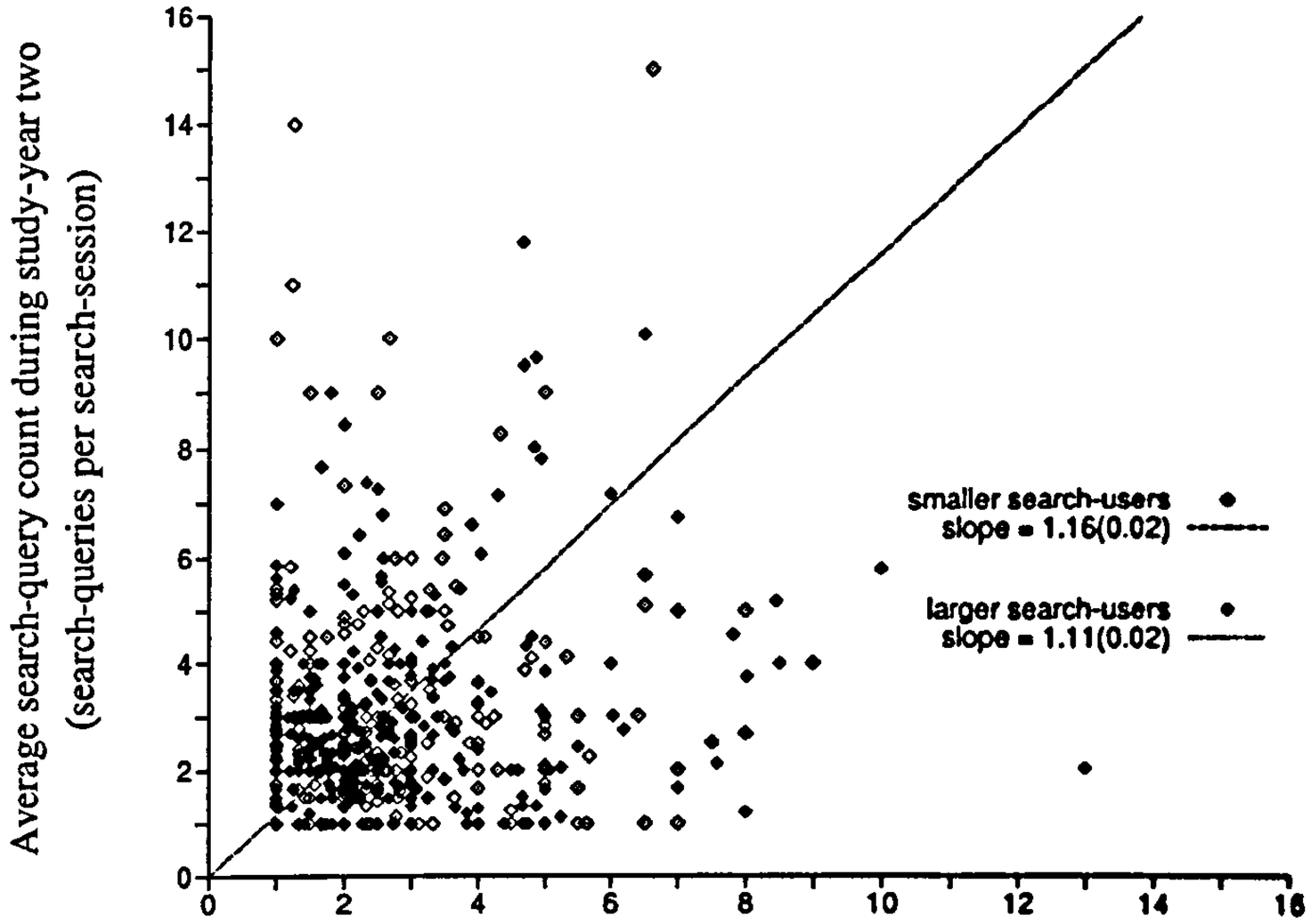

Figure E.35: Conditional distributions of AltaVista-Excite sample search-user's average search-query count by-joint-session-rate

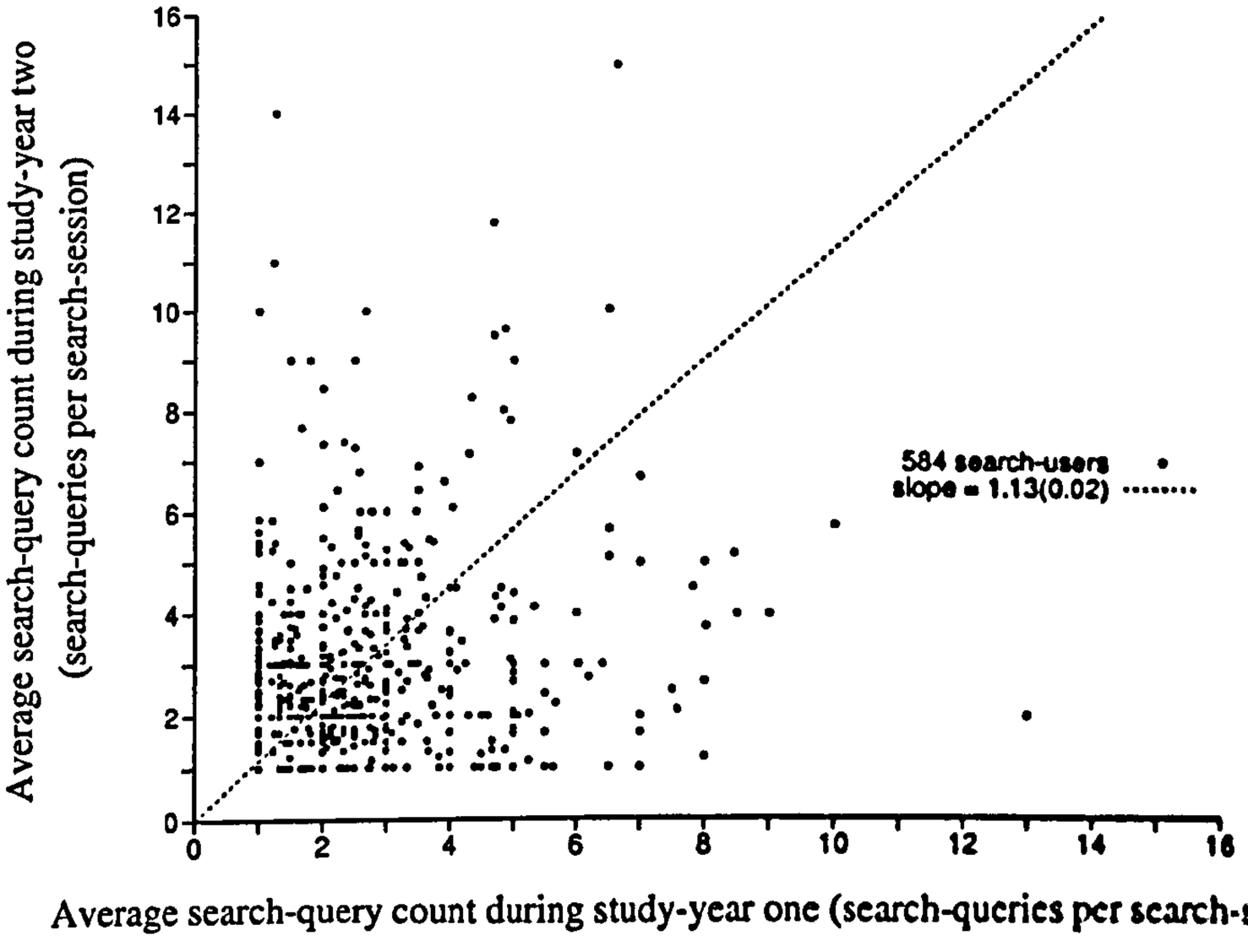

Figure E.36: Conditional distribution of AltaVista-Excite sample search-user's average search-query count

average search-term count

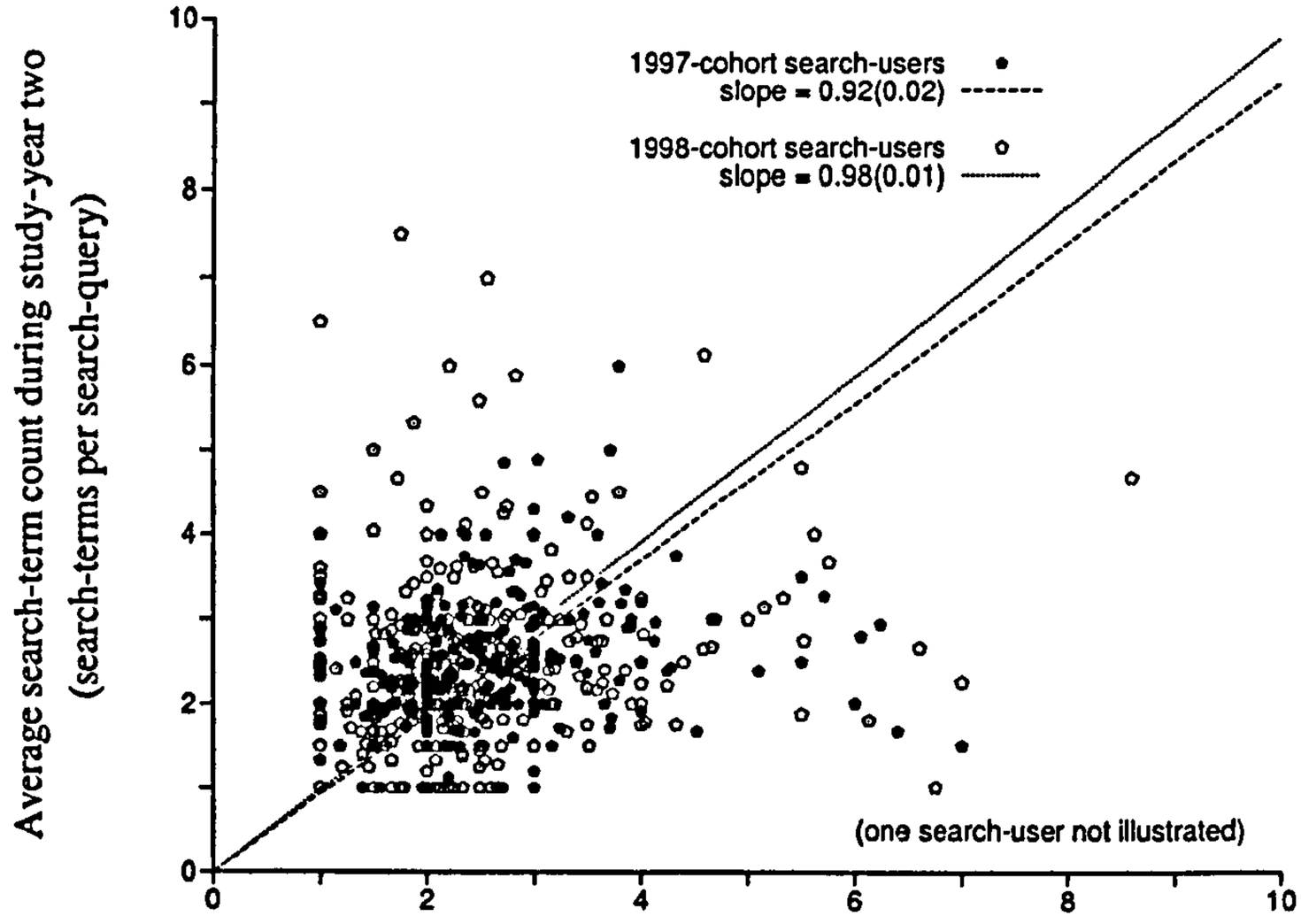

Figure E.37: Conditional distributions of AltaVista-Excite sample search-user's average search-term count by-cohort

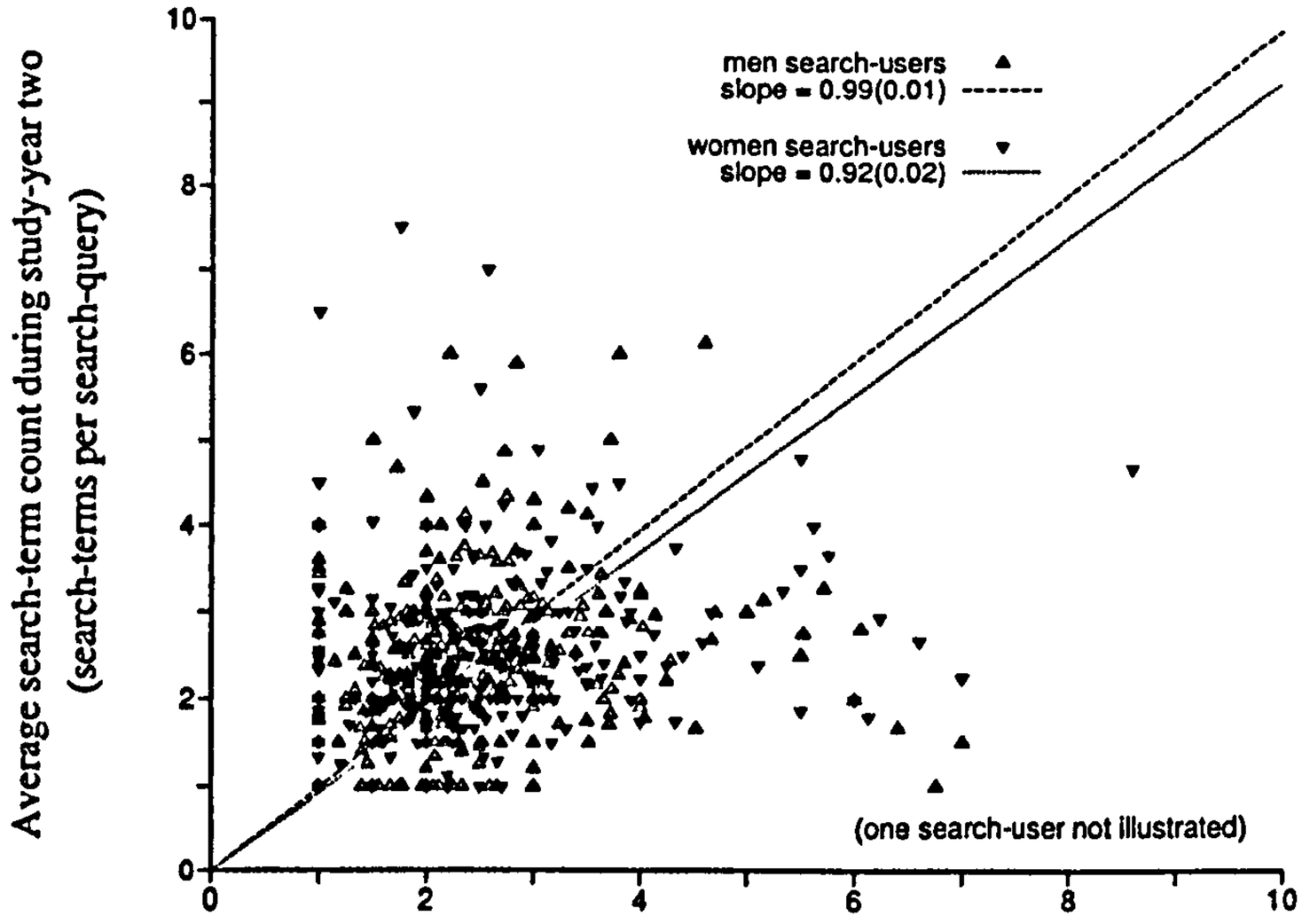

Figure E.38: Conditional distributions of AltaVista-Excite sample search-user's average search-term count by-gender

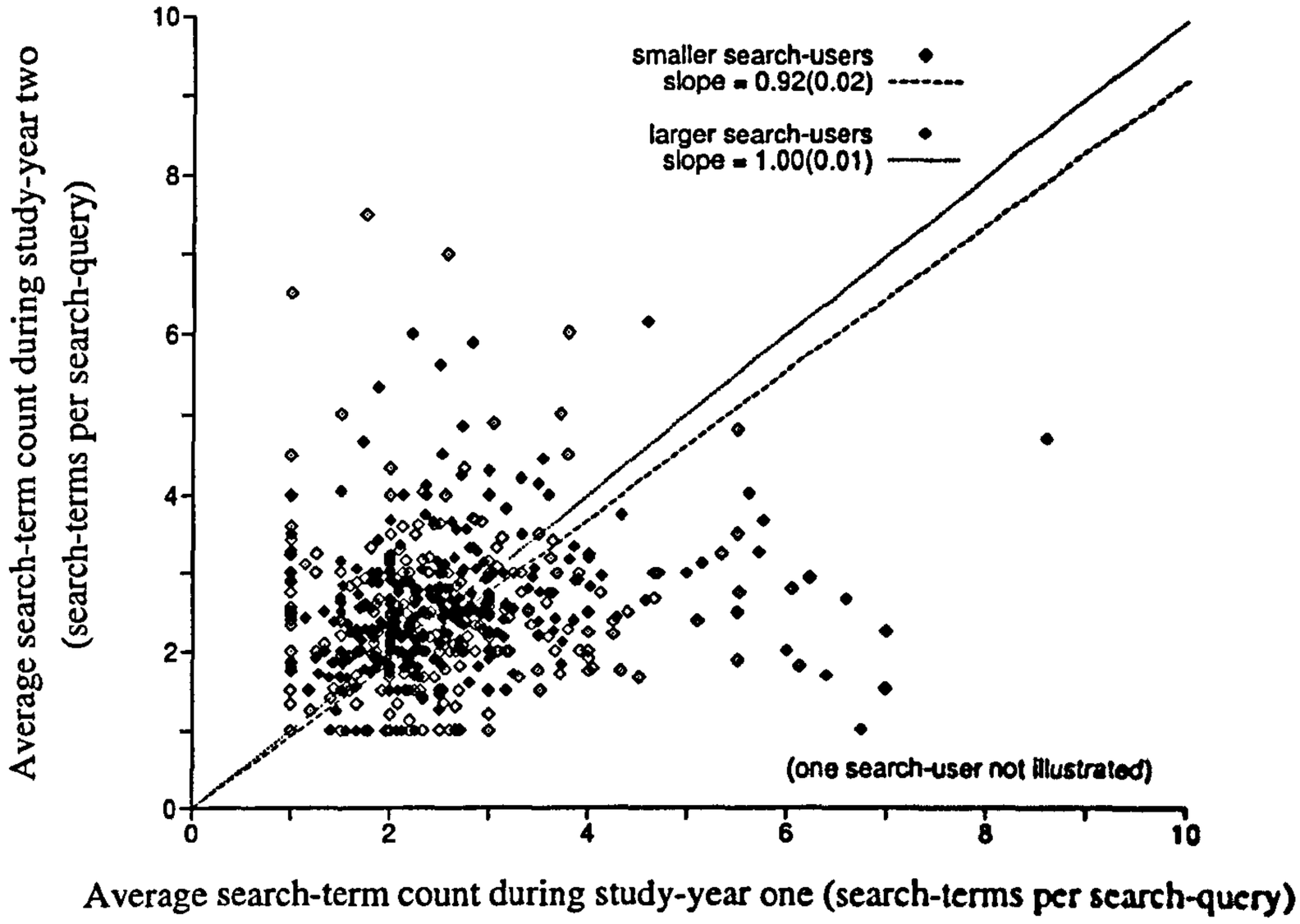

Figure E.39: Conditional distributions of AltaVista-Excite sample search-user's average search-term count by-joint-session-rate

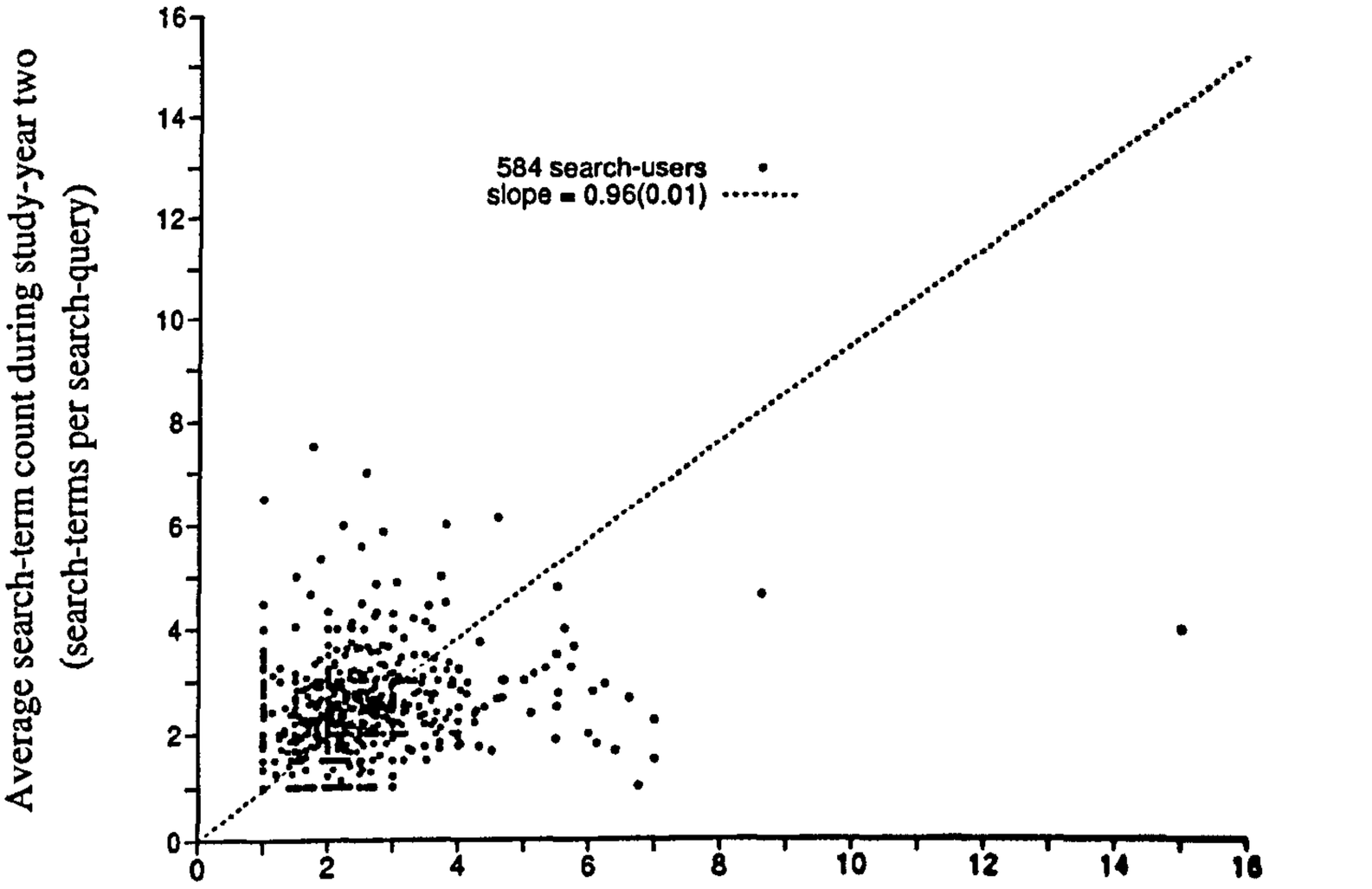

Figure E.40: Conditional distribution of AltaVista-Excite sample search-user's average search-term count